Journal of Applied Psychology

KENNETH E. CLARK, Editor
University of Colorado

Table of Contents

Donald G. Paterson: 1892–1961 351
The Assessment of Creativity in a Research Setting: William D. Buel and Virginia M. Bachner 353
Pilot Judgments of Simulated Collisions and Near Misses: A Comparison of Performance with Uncoded and Two-Tone Coded Models: John E. Robinson, Jr., Kenneth G. Cook, and Charles E. Zeleny 359
Variability of Stroke Width within Digits: Charles L. Hughes 364
The Interaction of Information Displays with Control System Dynamics in Continuous Tracking: R. W. Obermayer, W. F. Swartz, and F. A. Muckler 369
Cross-Validity of Procedures for Selecting Life Insurance Salesmen: Peter F. Merenda, Walter V. Clarke, and Charles E. Hall 376
Experimental Evaluation of Binary Codes for Console Display: Frank J. Minor and Stanley L. Revesman 381
Supervisory Procedures and Work-Team Productivity: J. S. Kidd and R. T. Christy 388
Bales' Interaction Process Analysis of Personnel Selection Interviews: Daniel Sydiaha 393
Protection against Impulse-Type Industrial Noise by Utilizing the Acoustic Reflex: James A. Chisman and J. Richard Simon 402
A Factor Analysis of the California Psychological Inventory: John O. Crites, Harold P. Bechtoldt, Leonard D. Goodstein, and Alfred B. Heilbrun, Jr. 408
Chance versus Nonchance Scores on the SVIB: Richard R. Stephenson 415
Visual Target Location as a Function of Number and Kind of Competing Signals: David M. Promisel 420
Effect of Personal Characteristics on Relationships between Attitudes and Job Performance: Francis D. Harding and Robert A. Bottenberg 428
The Stability of Guilford-Zimmerman Personality Measures: Jay M. Jackson 431
Relationships between Individual Proficiency and Team Performance and Efficiency: William M. Wiest, Lyman W. Porter, and Edwin E. Ghiselli 435

This is the last issue of Volume 45.
Volume Title Page and Contents appear herein.

American Psychological Association

Volume 45, Number 6 December 1961

Consulting Editors

ARTHUR BRAYFIELD, *Pennsylvania State University*
GEORGE E. BRIGGS, *Ohio State University*
NORMAN FREDERIKSEN, *Educational Testing Service*
LEONARD D. GOODSTEIN, *University of Iowa*
EDWIN R. HENRY, *Standard Oil Company of New Jersey*
JOHN HOLLAND, *National Merit Scholarship Corporation*
CLIFFORD E. JURGENSEN, *Minneapolis Gas Company*
LAURENCE S. MCGAUGHRAN, *University of Houston*
QUINN MCNEMAR, *Stanford University*
HAROLD F. ROTHE, *Fairbanks, Morse and Company*
THOMAS A. RYAN, *Cornell University*
A. C. WELCH, *University of New Mexico*

This journal gives primary consideration to original investigations in any field of applied psychology except clinical psychology, although a descriptive or theoretical article may be accepted if it represents a special contribution in an applied field. Quantitative investigations of interest or value to psychologists working in the following broad fields will be considered: vocational and educational prognosis, diagnosis, and guidance at the secondary and college level; personnel research in business, industry, and government; engineering psychology; industrial working conditions; research on opinion and morale factors; job analysis and classification research; market and advertising research.

Because of the large number of manuscripts submitted, authors should adhere to the rule of "brevity consistent with clarity." The typical manuscript should run to approximately 4,000 words. There is a lag of approximately twelve months between receipt and publication of an article. Authors may request advanced publication if they are prepared to pay the cost of printing the necessary extra pages.

Manuscripts should be addressed to the Editor:

Dr. Kenneth E. Clark
Office of the Dean
College of Arts and Sciences
University of Colorado
Boulder, Colorado

All manuscripts must be submitted in duplicate. Original figures are prepared for publication; duplicate figures may be photographs or pencil-drawn copies.

Manuscripts must conform to the style requirements described in the *Publication Manual of the American Psychological Association*.

Journal of Applied Psychology

Published bimonthly by the
American Psychological Association
Prince and Lemon Sts., Lancaster, Pa.
and 1333 Sixteenth Street N.W.
Washington 6, D. C.

$10.00 per volume $2.00 per issue

ARTHUR C. HOFFMAN
Production Manager

HELEN ORR
Promotion Manager

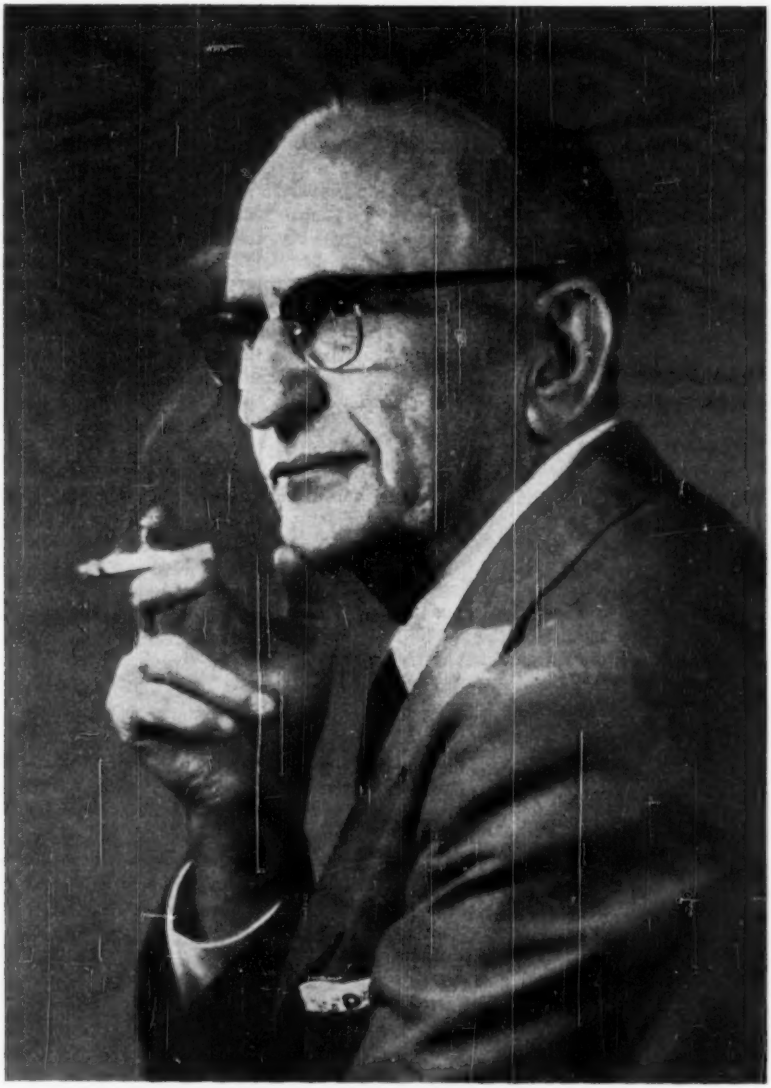
Subscriptions, orders, and business communications should be addressed to the American Psychological Association, 1333 Sixteenth St. N.W., Washington 6, D. C. Address changes must reach the subscription office by the 10th of the month to take effect the following month. Undelivered copies resulting from address changes will not be replaced; subscribers should notify the post office that they will guarantee second-class forwarding postage. Other claims for undelivered copies must be made within four months of publication.

Second class postage paid at Lancaster, Pennsylvania and at additional mailing places.

© 1961 by the American Psychological Association, Inc.

Journal of Applied Psychology

VOL. 45, NO. 6 DECEMBER 1961



DONALD G. PATERSON

DONALD G. PATERSON

1892–1961

Applied psychology lost one of its pioneers and one of the major contributors to its development with the death on October 4, 1961 of Donald G. Paterson. Returning to this country after participating in the fourteenth International Congress of Applied Psychology in Copenhagen, he was hospitalized in Minneapolis and died of cancer after a brief illness.

Paterson was Editor of this journal for 12 years, from 1943 through 1954, being the first person to perform this service under the APA's ownership and management of the journal. His list of publications exceeds 300 titles, including the authorship or co-authorship of *Minnesota Mechanical Ability Tests* (1930), *Physique and Intellect* (1930), *Men, Women, and Jobs* (1936), *Student Guidance Techniques* (1937), *How to Make Type Readable* (1940), *The Minnesota Occupational Rating Scales and Counseling Profile* (1941), *Local Labor Market Research* (1948), *Revised Minnesota Occupational Rating Scales* (1953), and *Studies in Individual Differences* (1961). It is characteristic of Paterson's mode of working and of his constant involvement with his many students that many of his books and articles were published jointly with students and colleagues. That his methods resulted in research productivity is demonstrated by the list of about 1,400 titles of articles published by his students which was prepared and published at the University of Minnesota at the time of Paterson's retirement in 1960.

Donald G. Paterson was born on January 18, 1892. He obtained his undergraduate and graduate training at the Ohio State University. Prior to World War I, he was an Instructor in psychology at the University of Kansas, where he met and married Margaret Young. His wife and their two children, Philip Paterson and Mrs. Robert C. Becker, survive him. From 1917 through 1919 he was Chief Psychological Examiner and a Captain in the Sanitary Corps of the United States Army. For the next 2 years he was a member of the Scott Company, the first of psychological consulting organizations.

Paterson moved to the University of Minnesota in 1921, and in 1923 was promoted to Professor. He retired in 1960 after 39 years of service. In this interval he became a leading figure in psychology nationally, in his own community, and in the University of Minnesota. He pioneered in the whole advance of student personnel work, vocational counseling, industrial and personnel psychology, and differential psychology. He played a major role in the 1930s and 1940s in the establishment of the University of Minnesota's Student Counseling Bureau, the Minnesota Employment Stabilization Research Institute, and the university's Industrial Relations Center. He was a founder and President of the American Association of Applied Psychology, and served as Secretary of the American Psychological Association for 6 years. He was a Diplomate in industrial psychology of the American Board of Examiners in Professional Psychology. He received an honorary LLD degree from Ohio State University in 1952, and in 1956 was selected to deliver the Walter V. Bingham lecture at his alma mater.

Since 1921, approximately 300 students have earned their MA degrees, and 88 students their PhD degrees with him as their major adviser. Additional thousands of undergraduate and graduate students have taken his courses in the psychology of individual differences, vocational and occupational psychology, and occupational counseling. Hundreds of his students, currently working in education, government, and industry, are applying their knowledge and engaging in research in areas of applied psychology. His imprint on applied psychology will continue to be seen in the work of these students, and in the contributions he has made to the knowledge of psychological principles and their applications in all areas of life.

Journal of Applied Psychology
1961, Vol. 45, No. 6, 353–358

THE ASSESSMENT OF CREATIVITY IN A RESEARCH SETTING

WILLIAM D. BUEL AND VIRGINIA M. BACHNER

Science Research Associates, Chicago, Illinois *Pure Oil Company, Palatine, Illinois*

Recently, Stein (1960a, 1960b), Sprecher (1959), Smith, Albright, Glennon, and Owens (1961), and Buel (1960) have presented both material and methods directed to the problems of the assessment and identification of creativity among research personnel in the industrial setting. These studies illustrate several methods of studying creative performance: environmental surveys, descriptive psychometrics, personal history information, and behavioral rating scale items validated for the purpose of differentiating between relatively more or less creative research personnel. The purpose of the study reported here was to investigate the descriptive and predictive validity of several psychometric instruments and a locally constructed forced-choice scale.

METHOD

Subjects

The subjects in this study were research personnel in the research center of a major cereal and animal feed company.[1] The subjects were selected with regard to whether their assignments were considered to be in an endeavor which called for creative behavior, regardless of whether they themselves were considered to be creative. In this manner 54 persons (48 males and 6 females) were selected for study.[2] The areas of technical endeavor represented by these persons were cereal, animal nutrition, and organic research with training in chemistry and biology as a common background. Fifteen held the doctor's degree, 4 held the master's degree, 28 held the bachelor's degree, and 18 others had training and experience in their field but did not hold a degree. The age of the group ranged from 22 to 59 with a mean of 34.9 years. Length of service with the research center ranged from 6 months to 27 years with a mean of 6.3 years.

Raters

Twenty-two raters evaluated the research personnel by means of the Forced-Choice Instrument. Twenty-one of these raters also served as ratees for some other rater. The number of ratees assigned to any rater ranged from 1 to 7. All 65 subjects were evaluated with the Forced-Choice Instrument. In addition, the codirectors of the research center participated in an over-all ranking of the 65 subjects on a criterion to be discussed below.

Psychometric Instruments

The psychometric instruments employed in this study were as follows:

1. Adaptability Test, Form A (12 persons) and Form B (42 persons). All 6 females responded to Form B. Forms A and B correlate .89 (Tiffin & Lawshe, 1954).
2. Thurstone Temperament Schedule, Form AH
3. Guilford-Martin Personnel Inventory
4. Allport-Vernon-Lindzey Study of Values, Revised Edition
5. Kuder Preference Record (Vocational—Form B1)
6. Forced-Choice Instrument

The items for the Forced-Choice Instrument were based on a previous study (Buel, 1960) in a petroleum research center. The items relate to research competence, methodology, technical application, personal approach to complexity, perseverance, energy level, synthesizing ability, uniqueness of approach, objectivity, personal confidence, use of symbolism and analogy, etc. Since the format of this instrument has not previously been described in the literature, it is briefly described here.

As previously indicated, the items were of a behavioral nature. The preference indices were in the form of the arithmetic means of the value (1 through 5) which had been assigned to each item by each rater at a time when those items were in check list form and were used to describe research personnel. The item discrimination indices were in the form of Pearsonian coefficients calculated against a hypothetical criterion of creativity, identical in all but organizational references to the one employed in this study.

The development of the tetrads for the scale was accomplished by matching two items on high preference value and two items on low preference value, one item in each pair having significant differentiating ability (.05 level) and one item in each pair having little or no differentiating ability. In addition to this form of matching, an attempt was made to match each pair of items on standard deviation (Buel, 1959). Finally, an attempt was made to match pairs of high and low preference value items on item length. In this way controls were imposed on central tendency, variability, and the desirability or attractiveness of an item as a function of length or brevity. The final

[1] The authors wish to thank Robert G. Coy and the other members of the Quaker Oats Company Research Center who made this study possible.

[2] A N of 65 is reported upon later in this paper for purposes of forced-choice validity and reliability.

forced-choice scale consisted of 13 tetrads, the items in each of which were to be ranked (1–4) in terms of their descriptiveness for the person being evaluated. Rundquist (1950) has recommended such a format, and it is at least reasonable to assume that the unpleasantness of forcing a choice is alleviated by this approach. The scoring of a scale in this form is achieved by scoring only Rank 1 and Rank 4, synonymous to scoring the most and least descriptive items.

Procedure

The testing of personnel was accomplished by a qualified psychologist between 6 and 9 months prior to the inception of this study. All persons tested were members of the research center staff at the time of testing and had not been selected for employment by means of these tests. The forced-choice evaluations and the criterion rankings were performed at the beginning of the data collection. All test and rating results were kept confidential.

Criteria

Three criteria of creativity were employed:

1. A hypothetical criterion of creativity stated as follows:

> You have been selected to be director of research for another cereal and feed company. The Quaker Oats Company Research Center has agreed to let you take with you to your new organization certain of the persons now at the Quaker Oats Company Research Center. You have been instructed by your new management to bring with you those persons who will make the most *significant, original,* and *lasting* contributions to research. Try to focalize your evaluation on the three underlined characteristics above, disregarding the individual's field of specialization.

The two codirectors of research evaluated each of the 65 persons on this basis. Each man was ranked in accordance with instructions which required that the group be divided into five subgroups, 13 persons per group, Group A being those most accurately described by the criterion, Group B those next best described, and so on down to Group E or those most poorly described by the criterion. Ranks 1 through 13 were then assigned to each member of each group. These letter-number combinations were later converted into one continuous ranking from 1 through 65, thus obviating the difficult task of continuously ranking 65 persons.

Employing the technique proposed by Hull (1922), each subject's rank score was converted into a linear scale score, amenable to treatment by linear correlation methods. The correlation between criterion raters was .73. Consequently, the average of each man's linear scale scores served as his criterion score.

2. Creativeness: a subsection of the company appraisal form, which was defined as follows:

> Ability to approach problems with an inquiring mind, with vision and imagination, to develop new ideas and improved methods, and to provide original solutions to problems.

Ratings on this criterion were in the form of values from 1 to 3 and were assigned as part of the yearly evaluation of personnel. The raters who assigned these values were in some instances the same persons who rated with the Forced-Choice Instrument. However, the 9–12 months between the two evaluations reduced any biasing effects as a function of memory of previous ratings.

3. Proportional Contribution to Patents Issued: wherein each single author patent was scored as one point; if authored by more than one person, the point was divided equally among the several authors.

The first criterion discussed was collected after the forced-choice evaluation of the participants, while the second criterion evaluation had been performed prior to the inception of the study. The third criterion was obtained from personnel records. Because of the temporal order in which the criteria were collected, relatively little bias is assumed to have crept in by way of slanted test scores or forced-choice evaluations.

Analysis

Tables 1, 2, and 3 present, respectively, the validities, intercorrelations between significant predictors, and criterion intercorrelations. One validity (.67), based on 65 cases, resulted from correlating the forced-choice scores with the first criterion (hypothetical), while with the same N for split-half calculations (7 tetrads vs. 6 tetrads), a reliability of .86 was obtained, increased to .92 by application of the Spearman-Brown formula (an approximation because of the uneven split). In all other situations N was equal to 54 or 42, as specified. Where necessary, adjustments were made in the ranks assigned by the codirectors and subsequently in the linear scale scores to account for the reduced number of cases. All correlations are in the form of Pearson product-moment coefficients.

Results and Discussion

Inspection of Table 1 leads to some interesting observations concerning the application of standardized psychometric material for discriminating between various levels of creative ability. Persons whose research contributions are *"significant, original,* and *lasting"* (Criterion 1) tend to be active dominant, and cooperative; are interested in literary pursuits; and are uninterested in clerical pursuits. Further, they are quite effectively identified by the Forced-Choice Instrument. Persons who are considered to possess "creativeness" (Cri-

TABLE 1

VALIDITY COEFFICIENTS

Predictor	Criterion 1 (Ranking) ($N = 54$)	Criterion 2 (Creativeness) ($N = 42$)[a]	Criterion 3 (Patents) ($N = 54$)
Adaptability Test	.22	.34**	.13
Thurstone Temperament Schedule:			
Active	.30**	.23	.14
Vigorous	.04	−.12	.01
Impulsive	−.09	−.07	−.03
Dominant	.26*	.22	.13
Stable	.04	21	−.08
Sociable	−.03	−.03	.11
Reflective	.22	.21	.05
Guilford-Martin Personnel Inventory:			
Objectivity	.20	.25	.08
Agreeableness	−.20	.18	−.07
Cooperativeness	.25*	.20	.06
Allport-Vernon-Lindzey Study of Values:			
Theoretical	.20	.22	.18
Economic	−.08	−.11	−.06
Aesthetic	−.07	.01	.07
Social	.02	.05	.05
Political	.08	.09	.01
Religious	−.12	−.24	−.20
Kuder Preference Record (Vocational):			
Mechanical	−.03	−.11	−.11
Computational	.11	.01	.15
Scientific	.17	.09	−.09
Persuasive	.11	.17	−.13
Artistic	−.22	−.02	.09
Literary	.29**	.29*	.24*
Musical	−.01	.04	.16
Social Service	.01	−.18	.03
Clerical	−.27*	−.19	−.16
Forced-Choice Instrument	.65***	.71***	.25*

[a] Unavailable data on Criterion 2 reduced N by 12.
* $p < .10$.
** $p < .05$.
*** $p < .001$.

terion 2) are more intelligent, are interested in things literary, and are quite easily identified by means of the Forced-Choice Instrument. Relative to the number of "patents issued" (Criterion 3), persons with literary interests as well as those identifiable by means of the Forced-Choice Instrument tend to patent.

Previously reported work lends credulity to some of the findings reported here. In relation to activity, Barron (1957) demonstrates that original persons tend to be more "energetic" whereas unoriginal persons tend to be "slow" and "apathetic." Elsewhere Barron (1955) hypothesizes and demonstrates that "Dominance" is characteristic of original persons.

TABLE 2

INTERCORRELATIONS AMONG SIGNIFICANT PREDICTORS
($N = 54$)

Predictor	Adaptability Test	Thurstone: Active	Thurstone: Dominant	Guilford-Martin: Cooperativeness	Kuder: Literary	Kuder: Clerical
Thurstone: Active	.13					
Thurstone: Dominant	−.05	.22				
Guilford-Martin: Cooperativeness	.25	.21	.06			
Kuder: Literary	.08	.16	.05	.02		
Kuder: Clerical	−.06	−.14	−.50	−.10	−.21	
Forced-Choice Instrument	.22	.17	.31	.23	.06	−.18

Maizell (1958), investigating the reading habits of research chemists, suggests that the more creative are given to greater amounts of technical reading on the job. Albright and Glennon (1961) find essentially the same tendency when differentiating between research on a "research" vs. "administrative" career aspiration criterion, while Roe (1952) describes eminent scientists as having developed a childhood habit of prodigious reading. Maschino (1959), considering the factors related to success in technical jobs, presents evidence to the effect that successful technical personnel are disinterested in clerical work and, as demonstrated in numerous other research reports, are significantly more intelligent. Cooperativeness, although not specifically corroborated by the work cited here, gains significance if one chooses to construe Ascendance (Barron, 1957) and Tolerance (Gough & Woodworth, 1960) as components thereof. Conversely, Roe's work relative to the researcher's need for personal independence (1952) and lack of personal relations skills (1956) would seem to be contradictory. The fact that the correlation of Cooperativeness with Criterion 3 (patents), a relatively objective criteria, is not significantly different from zero, may indicate that rater and subject bias in favor of a socially desirable characteristic like Cooperativeness has contributed to the magnitude of the correlations between it and two somewhat subjective criteria.

Table 2 presents the matrix of intercorrelations among the variables singled out for consideration as discriminators between relatively creative and noncreative research persons. It is apparent that, with the possible exception of Dominance and Clerical (−.50) and Dominance and the Forced-Choice Instrument (.31), the discriminating measures are independent of one another. Considering Active, Cooperativeness, Literary, and Clerical (Dominance contributes nothing because of high intercorrelations) as a predictive battery for Criterion 1 leads to a multiple correlation of .47 ($N = 54$), shrunken to .42 (Guilford, 1950; Kelley & Salisbury, 1926), significant at the .01 level of confidence. Selecting the Adaptability Test and Literary as predictors of Criterion 2 results in a multiple correlation of .43, shrunken to .41, significant at the .01 level ($N = 42$).

Although the authors have heretofore treated the Forced-Choice Instrument as a predictor of various criteria, it is equally justifiable to view it in the role of a criterion. Using the Forced-Choice Instrument ($N = 54$) as a criterion results in a validity for Dominance of .31 (.04 level) and for Theoretical of .27 (.06 level), correlating between themselves −.03 (presented here but not in tables). Combining the two eventuates in a multiple correlation of .41, shrunken to .39 (.01 level). It appears that, even within the limits of this study, the Forced-Choice Instrument is a reasonable criterion.

Table 3 presents the correlations between the criteria. Criterion 3 (patents) is more independent than Criterion 2 (creativeness),

TABLE 3

CRITERION INTERCORRELATIONS

Criterion	1	2
1 (Ranking) (N = 54)		
2 (Creativeness) (N = 42)	.64	
3 (Patents) (N = 54)	.44	.17

and Criterion 1 (hypothetical) is the least independent of the three, suggesting that it may be the most representative criteria. Criterion 3 is relatively independent but at the same time reflects the possibility that it is apparently not measuring creativity as conceived of by research directors and supervisors. The lower correlations between Criterion 3 (patents) and the other two criteria may reflect the many unaccounted for variables involved in securing patents or, perhaps, the entrepreneurship qualities operating in research personnel who obtain patents.

To use a patent criteria in such a study may be suspect when one recalls that certain of the subjects had been members of the research center staff a relatively short time. Obviously such persons, even if highly creative, had not had a chance to patent within this organization. Patents were included as a criteria more because of availability than because of their purity as a measure of creative ability, even had all persons been staff members long enough to make such a criterion equitable. The authors feel that mere profusity of patents is not a true indicator of creative ability.

Future investigations might develop the merit of patents as criteria if the differential values of patents were considered. If some method of quantifying the value of individual patents (perhaps by a panel consisting of the patent author and several experts) were developed, a more meaningful patent criterion might result. With regard to the results of this study, subtests which approach significance for one or several criteria, but which have not previously been discussed, might bear further scrutiny and item analysis. Stable, Reflective, Objective, Agreeableness (negative), Artistic (negative), and Religious (negative) might well be developed into significant indicators of the personality and value structure of creative persons.

SUMMARY

Psychometric instruments of varying nature were used to evaluate research personnel in a major cereal and feed company research center. Three criteria of creativity were collected against which the psychometric instrument validities were calculated. Certain intelligence, interest, and personality variables and the Forced-Choice Instrument were demonstrated to be valid, singly and in combination, for differentiating between relatively creative and noncreative research workers.

REFERENCES

ALBRIGHT, L. E., & GLENNON, J. R. Personal history correlates of physical scientists' career aspirations. *J. appl. Psychol.*, 1961, **45**, 281–284.

BARRON, F. The disposition toward originality. *J. abnorm. soc. Psychol.*, 1955, **51**, 478–485.

BARRON, F. Originality in relation to personality and intellect. *J. Pers.*, 1957, **25**, 730–742.

BUEL, W. D. Stability of preference indices in forced-choice rating scale items. *Engng. industr. Psychol.*, 1959, **1**, 134–137.

BUEL, W. D. The validity of behavioral rating scale items for the assessment of individual creativity. *J. appl. Psychol.*, 1960, **44**, 407–412.

GOUGH, H. G., & WOODWORTH, D. G. Stylistic variations among professional research scientists. *J. Psychol.*, 1960, **49**, 87–98.

GUILFORD, J. P. *Fundamental statistics in psychology and education*. New York: McGraw-Hill, 1950.

HULL, C. L. The computation of Pearson's r from ranked data. *J. appl. Psychol.*, 1922, **6**, 385–390.

KELLEY, T. L., & SALISBURY, F. S. An iteration method for determining multiple correlation constants. *J. Amer. Statist. Ass.*, 1926, **21**, 282–292.

MAIZELL, R. E. The most creative chemist reads more. *Industr. engng. Chem.*, 1958, **50**, 64A–65A.

MASCHINO, A. Factors related to success in technical jobs. Midland, Mich.: Dow Chemical, 1959. (Mimeo)

ROE, A. A psychologist examines sixty-four eminent scientists. *Scient. Amer.*, 1952, **187**, 21–25.

ROE, A. *The psychology of occupations*. New York: Wiley, 1956.

RUNDQUIST, E. A. Personality tests and prediction. In D. H. Fryer & E. R. Henry (Eds.), *Handbook of applied psychology*. New York: Rinehart, 1950. Pp. 182–191.

SMITH, W. J., ALBRIGHT, L. E., GLENNON, J. R., & OWENS, W. A. The prediction of research competence and creativity from personal history. *J. appl. Psychol.*, 1961, **45**, 59–62.

SPRECHER, T. B. A study of engineers' criteria for creativity. *J. appl. Psychol.*, 1959, **43**, 141–148.

STEIN, M. I. *The research environment survey*. Chicago: Science Research Associates, 1960. (a)

STEIN, M. I. *The survey for administrators*. Chicago: Science Research Associates, 1960. (b)

TIFFIN, J., & LAWSHE, C. H. *Examiner manual for the Adaptability Test*. Chicago: Science Research Associates, 1954.

(Received December 9, 1960)

Journal of Applied Psychology
1961, Vol. 45, No. 6, 359–363

PILOT JUDGMENTS OF SIMULATED COLLISIONS AND NEAR MISSES:

A COMPARISON OF PERFORMANCE WITH UNCODED AND TWO-TONE CODED MODELS [2]

JOHN E. ROBINSON, JR., KENNETH G. COOK, AND CHARLES E. ZELENY

Applied Psychology Corporation

Aircraft pilots agree that avoiding midair collisions in Visual Flight Rules weather [1] involves more than merely sighting the intruder aircraft; but they do not agree on just how much more: how many kinds of information are needed and what degrees of precision are required (Calvert, 1958; Robinson, 1959). This question continues to arouse arguments in ready-rooms, at manufacturers' conferences, and at flight safety meetings and regulatory deliberations.

Projector and Robinson (1958) in an analytical survey of aircraft navigation light systems list eight kinds of information mentioned in aviation literature as important, to some unspecified degree, in avoiding collisions during "good" weather conditions (Visual Flight Rules):

Presence or location	Heading or course
Identification	Airspeed
Attitude	Distance or range
Relative altitude	Intended maneuver

Presumably, it was felt that if a pilot could get these kinds of information about "intruder" aircraft, he would be able to make faster, more accurate decisions about the possibility of collision, and thus, whenever avoidance was necessary, make the correct maneuver sooner and more accurately.

A research approach to the evaluation of this basic concept would seem to begin with the measurement of visual perception of an object moving through three-dimensional space. A first task would be to measure the speed and accuracy of visual perception of simple or elemental objects. Subsequently, perception of objects with more complex visual characteristics would be measured, after additional items of visual information had been provided.

[1] Part 60, Civil Air Regulations, with amendments.

Aircraft shape has been shown to be advantageous in the perception of relative motion when compared to the relative motion of a spherical object (Kelley, Bowen, DeGroot, Frank, & Channel, 1959). Individual observers were seated in a corner of a black curtained "room." A visible stimulus object whose support was invisible was moved across the room in a random sequence of straight-line directions. The observer had to indicate his perception of the relative motion of the object. His answers were scored on the basis of speed and accuracy. The first object was a small, silver painted sphere. Very little visual information was available from the shape of the stimulus, and the scores reflected certain amounts of erroneous relative-motion judgment. It was reported that when the sphere was replaced with a small, silver painted aircraft model, judgments of relative motion were made significantly faster and more accurately.

The experiment reported in the present text [2] was designed to investigate and provide the answer to the next logical step in this research progression: Will the relative motion of aircraft models painted half dark and half light be perceived sooner and more accurately than the relative motion of one painted all light?

The answer was sought through visual presentation of a number of simulated collisions and near-collisions, using an F-100 flight simulator and an F-151 aerial gunnery trainer attachment. Pilots who served as subjects used apparatus which made continuous records of their judgments during the simulated

[2] Part of research under Contract No. FAA/BRD-127, Analysis and Comparison of Visual Collision-Avoidance Systems, funded by the Bureau of Research and Development (now, Aviation Research and Development Service), Federal Aviation Agency.

flight situations. The three judgments possible were "Undecided," "Collision," and "Miss."

The paint patterns used on the intruder aircraft models were in no way intended to presage operational paint schemes for full-sized aircraft. Rather, they were designed to code certain visual aspects of the aircraft to see if such coding would help pilots make better judgments than could be made with an uncoded model. If the aspect information had turned out to be of value, application would have been developed for both daytime operations (paint) and nighttime operations (lights).

Procedure

Each pilot who participated in this study solved his test problems while seated in the cockpit of an F-100 jet fighter in one of the simulator rooms at the Federal Aviation Agency's National Aviation Facilities Experimental Center, Atlantic City, New Jersey. In front of this cockpit was a concave plaster screen, approximately two stories in height and extending 240 degrees in azimuth. The general level of lighting represented moonlight conditions (7 to 10 foot-lamberts ambient light). No terrain features were visible, but a fuzzy horizon was distinguishable at the border of the "moonlit" area.

A closed circuit television system was used to project an image of a scale model B-47 aircraft in a series of different courses relative to the course of the cockpit. These included flight towards the F-100 and away from it on one of six relative headings. Relative altitude of the courses was also varied, with the B-47 image being made at various times to appear to be ascending towards, descending upon, or flying level with the F-100. The outcome was sometimes a "collision," at other times a "near-collision," and at others a clear miss.

The "intruder" image appeared on the screen as a white area with a brightness of .55 foot-lamberts. The brightness of the "sky" (the area of the screen above the horizon line) was .1 foot-lambert. The maximum simulated range was 24,000 feet, at which distance the television image of the B-47 appeared only as a white dot. At distances closer than the minimum possible range (600 feet) the intruder image disappeared.

A separately located television camera was focused on the scale-model aircraft, which by means of a mount could be moved various distances from the camera to simulate changes in the intruder's distance. A gimballed suspension system permitted maneuver of the scale model in pitch, roll, and yaw. The amount and speed of rotation in each of these three flight dimensions, or any combination of them, were controlled by analog computer equipment; complex driving mechanisms assured that the apparent motion and maneuvers of the intruder image were faithfully maintained with respect to the cockpit in which the pilot sat.

The fact that computer settings could be made to start a predetermined flight situation, and carry it to completion, made it possible to have standardized test problems. Thus, the selected flight situations could be presented repeatedly with reasonable assurance that successive presentations were identical.

Six FAA pilots (two military, four civilian) served as subjects. All had had flight experience greatly in excess of the 1,000-hour minimum which the investigators established. All were actively engaged in flying at the time of the experiment, and the nature of their duties at the experimental center supports the belief that all possessed more than average interest in experimentation aimed at improving air safety.

Each subject was provided with 2 hours of orientation and practice prior to participating in four test sessions. Each test session contained 6 practice problems and 12 test problems involving one model. The order in which the models were assigned to test sessions as targets was counterbalanced to prevent bias from learning or practice effects. With individual problem times varying from 34 to 72 seconds, plus the time required to reset the computer for each new problem, single test sessions lasted approximately 50 minutes.

As soon as a problem was ready for presentation, the experimenter alerted the pilot ("Your next problem will appear soon at approximately 10 o'clock and level"). The pilot was not required in any case to make evasive maneuvers, but instead was asked only to judge whether or not the B-47 image would "collide" with his aircraft, assuming that speeds and flight paths would remain constant.

By pressing one of three keys on a board strapped to his knee, the pilot indicated either that he was or was not on a collision course, or that he was undecided. The keys were wired to a pen recorder in the control room. The resulting pen-trace record provided a means of measuring the time and nature of each judgment. A second pen, wired in circuit with the master switch, recorded the exact starting time of each problem.

The intruding B-47 images were visually coded in four separate patterns. One of the scale models (Model A) was painted entirely white, therefore requiring the pilot to make his collision judgments only on the basis of apparent aircraft structure and relative motion. The other three models were painted in two-tone schemes, as follows:

Model B:	front half	white
	back half	dark
Model C:	top half	white
	bottom half	dark
Model D:	left half	white
	right half	dark

The complete experiment consisted of two presentations each of 36 test problems for each of the four intruder models. The total number of problems available for analysis from the six pilots thus came

to 288. (The 144 practice problems were not included in the analysis of results.)

Pilot judgments were scored in terms of accuracy and speed. Accuracy scoring consisted of "correct" and "incorrect" categories, but were further classified in terms of initial judgments (decisions made earliest in each problem) and final judgments (the decision that was being indicated at the end of each problem).

For scoring of speed, five categories were delineated: (*a*) "Search" time (from start of problem to beginning of Undecided trace), (*b*) "Undecided" time (total time the pen trace was in Undecided area), (*c*) time to reach initial collision judgment (the sum of Search and Undecided times), (*d*) "Correct Response" time (total time the pen trace was indicating the correct problem outcome), and (*e*) "Incorrect Response" time (total time the pen trace was indicating an incorrect problem outcome). Because there were different time lengths among the 36 problems, actual time values were transformed into proportions of total problem time, and analysis was based on the proportional values.

After all the pilots had completed all test sessions, individual interviews were conducted to obtain information relevant to the data that had been collected. Topics included in the Interview Guide were: adequacy of training, convenience of response apparatus, realism of the visual presentations, nature of paint-coding that yielded best judgments, chances of correctly anticipating particular flight situations, appropriateness of test session length, and bases on which judgments were made.

TABLE 1

ACCURACY OF PILOTS' JUDGMENTS

Pilot	Model paint-coding: White	Front-back	Top-bottom	Left-right	All models
	Initial judgments				
1	6	7	6	3	22
2	5	6	9	6	26
3	8	9	12	8	37
4	11	11	8	9	39
5	12	10	8	10	40
6	9	11	7	10	37
All pilots	51	54	50	46	201
	Final judgments				
1	10	12	11	12	45
2	10	11	10	10	41
3	10	10	12	12	44
4	12	11	9	10	42
5	12	12	11	12	47
6	10	10	11	11	42
All pilots	64	66	64	67	261

Note.—The maximum possible values are as follows: single cell, 12; "all models" cell, 48; "all pilots" cell, 72.

RESULTS

The analysis of the data resulting from the experiments can be discussed conveniently under three topics: accuracy of judgments, speed of judgments, and effect of flight situation.

Table 1 summarizes the results of scoring the pilot judgments for accuracy according to individual pilots and individual target paint schemes. The numbers of correct judgments, both initial and final, indicate that the pilots were highly skilled. Of the 288 initial judgments, 201 (70%) were correct; while 261 (91%) of final judgments were right. The subjects, as individuals, varied widely in accuracy of initial judgments, ranging from 46% to 83% correct. On final responses, however, the six pilots were very nearly equally correct, ranging from 85% to 98%. The variation in response accuracy classified according to the four paint-coded models is small, both for initial and final judgments. Correct initial judgments ranged between 64% and 75%, while on final judgments the correct responses ranged from 89% to 93% of the 288 problems.

Examination of the results of scoring the pilot judgment records for time of decision reveals an interesting pattern of variability. Four of the six pilots required nearly identical amounts of time to detect the targets, while two pilots were markedly faster than the others. When Search time and Undecided time are summed, a considerable spread in proportion of problem time (23–69%) was evident among the six pilots. The variation in the proportion of time spent in making Correct judgment responses ranged from 27% to 54%; the time spent in making Incorrect responses varied from 4% to 26% of total problem time. Only in the case of Undecided time was the variability among pilots statistically significant beyond the .01 level. The different paint-codings did not cause statistically significant variation in the proportion of total

problem time spent on any response category except for Search time. Here, all of the two-tone models show longer times than the all-white model. This undoubtedly resulted from the fact that the "dark" portion of the two-tone models (which was in actuality a mottled gray) reduced the total brightness of these images as compared with the brightness of the all-white image.

Examination of scores to determine the effect of relative flight path on the accuracy and speed of pilots' judgments reveals that this variable did not appear to affect the responses significantly. Although individual differences occur among pilots and among categories of data, the variability does not reach statistical significance at conventional levels. Consideration of the variability in practical or operational terms similarly shows no difference among responses to the several targets that would warrant support of the two-tone codings as beneficial to the visual evaluation of collision threats.

The amount of vertical separation in flight path (at the completion of the problems) appears to have a general relation to the speed of pilots' judgments. That is, examination of the Undecided time reveals that the greater the vertical miss, the shorter the indecision. The closest vertical miss (400 feet) and both of the horizontal crossover misses required the longest times for making decisions.

In the postexperiment interviews, most of the pilots felt they did best with the all-white model, and they felt that the top-bottom coded model was probably most difficult. Their comments about the orientation and training were all favorable, and indicated that a little less training might have been adequate. The response apparatus was judged convenient to use and free from ambiguity. The length of the test sessions appeared to produce no fatigue or feelings of unreasonable demands.

Comment on the realism of the problems pointed up the fact that seeing a bright aircraft against a dark sky—as if a distant spotlight were illuminating only the intruder—would be somewhat rare in nature. Nevertheless, the pilots stated that the appearance of the problems was sufficiently realistic to permit their use in this phase of collision avoidance research.

In reaching their collision judgment decisions, all pilots reported that their basic method was to look for relative motion of the intruder image. If the image appeared to have any perceptible movement, the chances were good that a miss would result. Since the cockpit was perfectly steady during the problems, the relative motion cue could be (and was) utilized to great advantage.

Conclusions

It is felt that two major conclusions may be made from the data obtained in this investigation.

First, since the accuracy of collision judgments of visually coded targets showed no statistically significant improvement over judgments made with an uncoded target, sector coding of the type used here does not appear to be a substantial aid to the pilot who must make a visual evaluation of midair collision threats.

Second, the consistent reports of the pilots in the postexperiment interviews indicate that their fundamental visual cue for making collision-or-miss decisions is perceived relative motion of the target image.

The investigators believe both findings have important implications in determining the amount and kinds of information a pilot requires to avoid midair collisions in VFR operations.

Summary

An F-100 flight simulator and an F-151 aerial gunnery trainer with a black-and-white, closed circuit television system were used to present standardized visual flight problems involving a variety of relative course situations and altitude changes, resulting sometimes in "collisions" and sometimes in one of a variety of "near-collisions."

Six experienced pilots completed a total of 144 training and practice problems and 288 test problems. Objective records taken during each problem permitted quantitative measurement of the accuracy and speed with which the pilots were able to identify the outcome as "collision" or "miss."

Four target images, all generated from miniature B-47 models, were presented in counterbalanced test session order and in random sequences of problems. One model was painted all white, the other three were treated with dark and light paint to identify pairs of the cardinal aspects of any aircraft: top versus bottom, front versus back, and left versus right.

The objective of the investigation was to determine whether or not these visual codings would improve pilot ability to judge the outcomes of the simulated flight situations.

Statistical analysis was accomplished in terms of descriptive values for accuracy and speed, and of inferences drawn from analysis of variance. No statistically significant differences relevant to the objectives of the study were obtained between the two-tone coded models and the all-white uncoded model. The investigators believe it is appropriate to conclude that visual coding of the type studied is not of significant aid to a pilot in visually determining the existence of a collision threat or a safe passage; and, that pilots rely on perception of relative motion more than any other factor in judging the probability of colliding with another aircraft.

REFERENCES

CALVERT, E. S. Some operational aspects of the use of aircrew-interpreted devices for preventing collision in the air. (International Civil Aviation Organization, Working Paper AIR C-WP/49, 17 July 1958.) Presented to the Second Meeting of the Airworthiness Committee, Montreal, Canada, July 3, 1958.

KELLEY, C. R., BOWEN, H. M., DEGROOT, S. G., FRANK, P. K., & CHANNEL, R. C. Relative motion: II. The nature of relative motion situations. *USN Train. Dev. Cent. tech. Rep.*, 1959, No. 316-1.

PROJECTOR, T. H., & ROBINSON, J. E., JR. *Midair collision avoidance with navigation light systems.* Washington, D. C.: Airways Modernization Board, 1958.

ROBINSON, J. E., JR. *Human engineering tests of selected aircraft anti-collision light systems.* Washington, D. C.: United States Navy Bureau of Aeronautics, 1959.

(Early publication received April 24, 1961)

Journal of Applied Psychology
1961, Vol. 45, No. 6, 364-368

VARIABILITY OF STROKE WIDTH WITHIN DIGITS

CHARLES L. HUGHES

International Business Machines, New York

Typically, studies of stroke width of numerals have used a uniform stroke width overall. Chapanis, Garner, and Morgan (1949) suggested that this may not be the optimum condition, and they outlined some further research needed in which the various components of the digits are emphasized by varied stroke width within the numeral. This offered a new dimension in digit design. Studies directly relating to the present problem are in short supply, with the exception of an article by Soar (1958). Starting from an analysis of research results showing the confusion resulting from the use of conventional arabic numerals, Soar developed new sets of arabic digits in which parts of the numerals were distorted in form and/or stroke width to emphasize a difference between them. The common elements were minimized and the unique elements were emphasized. Six of the experimental digits proved to be significantly more visible than the usual arabic digits.

The relative readability of different digits depends on the criterion used as a legibility measure, the style or form of the numerals, and the stroke width. The digits 0, 3, 6, and 9 are most often confused with each other. They usually have curved rounded outlines. In most studies the digit 7 generally has the best visibility. This gives part of the rationale for the trend toward angles and away from curves. "Angularity" is the term applied to the de-emphasis of roundness or curvature of form and the use of sharper curves and straight lines in the numeral, as in the 0, 3, 6, and 9 versus the 7.

Two earlier experiments by Soar (1955a, 1955b) served to fill in the background leading up to the present problem. Significant interaction between stroke width, illumination level, and figure-ground contrast was demonstrated, with stroke width to height ratios the same overall (constant even stroke over the entire figure). Conventional arabic numerals were used, but modified in form somewhat to more angularity.

Additional evidence has been compiled which suggests that angular straight-line figures are more readable with symbolic digits receiving increasing but contradictory attention. Symbolic digits are simplified straight-line figures that have cognates in the arabic numerals. They have no curved lines. Some figures that are in between in form (distorted, angular digits) were used by Berger (1944a, 1944b).

In an often cited study Mackworth (1944) devised some new numerals and letters that simplified the components and curvature of conventional forms. The new angular digits were readable at significantly greater distances than the conventional ones. Several other studies have made comparisons of the relative visibilities of different designs, namely, Brown, Lowery, and Willis (1951) and Lansdell (1954). McLaughlin (1948) compared some standard sets of digits against new designs having variability of form and stroke width within the numerals. The results show that the form of the digits was not significant, but the within-varying stroke width was. This is at odds with the previously reported studies where form and even-overall stroke widths were both significant. The results are further confused by a retest of McLaughlin's study (Soar, 1958) in which the digit forms were significant. Varying stroke widths within conventional numerals complicates the problem so that the curvature and stroke width changes cannot be separated.

Many studies have been done on the effect of stroke width to height ratios upon visibility. In all cases the same stroke width was used for the entire set of digits and was not varied within any set or within any single digit. Berger (1944a, 1944b), Brown (1953), Brown and Lowery 1949) all obtained similar results. In addition Shapiro (1951), Loucks (1944), Craik (1941), Kuntz and Sleight (1950), Forbes and Holmes (1939) all obtained similar results with optimal stroke width to height ratios falling between 1:6

and 1:8. Actually only McLaughlin (1948) and Soar (1958) have varied stroke widths within digits. Furthermore, as we have mentioned they could only make a partial investigation of the relative contributions of stroke widths and form variation to readability of digits.

All these studies have included variations of conventional arabic numerals. Cohen and Webb (1953) suggested symbolic numerals made from a six-element straight-line matrix. These particular symbolic digits did not give as good a performance as conventional numerals. The study did result in a suggested eight-element straight-line matrix which was tested (Allusi & Martin, 1958; Allusi & Muller, 1956). Performance on these numerals was not very different from performance on conventional numbers. The symbolic digits, however, offer a way to approach the problem of form, within-variable stroke width and their interaction.

The primary conclusions which may be drawn from the previous research on digit design cited above would appear to be as follows:

1. The degrees of boldness of stroke lie within the range (SW/H ratios) 1:6 to 1:8 for best visibility.
2. Overall width to height ratios may range from 1:1 to 7:10.
3. Angularity increases visibility.
4. Increased inclosed white space aids visibility.
5. No adequate theory or hypothesis has yet been formulated to account for the differences in visibility demonstrated in research studies.

Questions which remain to be answered are:

1. What are the relative visibilities of contentional arabic numerals and symbolic numerals?
2. Do form and stroke width interact in digit perception?
3. Do some digits require greater or lesser stroke width for best visibility than other digits (should stroke width vary within a set of 10 digits)?
4. Do some digits require a variation of stroke width within that particular digit, through stroke width emphasis of some part of the digit for best visibility (should stroke width vary within a single digit)?

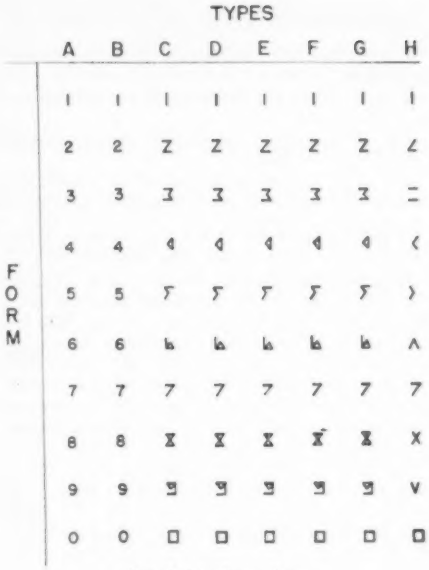


FIG. 1. Configurations.

PROBLEM AND METHOD

The present study was an attempt to study systematically the effect of variations in within-variable stroke width, conventional versus symbolic digits, emphasizing the unique aspects of digits through stroke boldness, angularity versus curvature, and the interaction of these variables.

Subjects. The subjects (*Ss*) were all male students enrolled in a course in Engineering Psychology, all with normal vision or with optical correction.

Apparatus. The stimuli were a set of cards on which the digits were drawn, one digit per card, each digit $\frac{3}{4}$" high and were of two stroke widths. The stroke width to height ratios (SW/H) were either 1:6 or 1:8. These stroke widths then were either *emphasized* (1:6) or *unemphasized* (1:8). Eight "types" of digits, 10 of each (1 through 0) were used for a total of 80. Some, such as the 1 and 0, were duplicates of other types and were included to prevent inbalance in the frequency of presentation of each form and as an offset for guessing habits on the subjects. The types were (see Figure 1):

Type A—conventional arabic, even-overall stroke width, SW/H 1:8

Type B—conventional arabic, even-overall stroke width, SW/H 1:6

Type C—symbolic, even-overall stroke width, SW/H 1:8

Type D—symbolic, even-overall stroke width, SW/H 1:6

Type E—symbolic, emphasized horizontal components, SW/H 1:6 and 1:8

Type F—symbolic, emphasized vertical components, SW/H 1:6 and 1:8

Type G—symbolic, emphasized unique components, SW/H 1:6 and 1:8

Type H—symbolic, special design with only unique components, the common elements absent, SW/H 1:6

Training. Two weeks and one week prior to the data gathering the *S*s were familiarized with both the conventional and symbolic digits. Specific training using the actual stimulus cards were given to all *S*s and the method used in the design of the digits was fully described. The level of learning achieved was tested by having all the *S*s individually reproduce accurately all of the configurations used in the study. The symbolic digits have cognates in the conventional arabic numerals, facilitating positive transfer of training.

Procedure. *S*s were given individual trials, and a complete trial for an *S* consisted of one round through the entire set of stimulus cards. The cards were reassembled in random order for each *S*'s trial. The initial stimulus to *S* distance was 21 feet, and this was decreased by one-foot intervals until the *S* could read the digit correctly. The stimulus cards were illuminated at 100 foot-candles with all of the test room conditions closely following optimum visual conditions. The *S*s were given instructions not to give a response until they were reasonably confident or could see the digit well. Wrong responses were not scored, and the experimenter (*E*) gave no reply to errors; however, the instructions were repeated. Only scattered guesses occurred and not more than twice by any one *S* so there was no evidence to show any *S* was using an elimination process to guess responses by overserving *E*'s behavior on wrong responses. After a correct response, the *S* returned to the starting position and the next stimulus card in the series was presented. This routine continued the same for the entire set of stimuli. Data were recorded in terms of distance in feet at which the *S* read the stimulus card correctly.

Statistical Analysis. The effects of form, stroke width, and the interaction were analyzed by factorial analysis of variance (for just the complete symbolic digits, Types C through G, omitting the conventional numerals and the special incomplete symbolic designs). The analyses of all types were done by 10 separate treatment × subjects analysis of variance tests, one for each form, 1 through 0. In addition 10 Duncan's shortest significant range tests were done to indicate which particular means were significantly different from which others.

Definition of terms. Throughout this report the following terms will be used in accordance with the definitions given here:

Form—the shape or contour of the digit irrespective of the stroke variations (e.g., the 2 has one form and the 3 has a different form, etc.)

Stroke Width—the ratios of SW or line emphasis (e.g., 1:6 bold stroke, emphasized; 1:8, unemphasized, nonbold stroke)

Configuration—the total or whole digit including form and SW variations (e.g., the entire stimulus complex, the total stimulus)

Component—any part or stroke or contour of a digit (e.g., the stimulus elements, the diagonal of the 7, the lower loop of the conventional 5, etc.)

Uniqueness—the condition of having components or element combinations that are not present in other configurations, either through changes in SW or form (e.g., only the 7 has a top horizontal bar and a diagonal stroke in combination, or combinations of stroke width variations and arrangement of lines)

TABLE 1

RELATIVE EFFECTS OF FORM AND STROKE WIDTH

Factorial analysis of variance

Source	*df*	*SS*	*MS*	*F*
Form	9	990	110.00	—
Stroke Width	4	40	10.00	—
F × SW	36	373	10.36	2.89*
Subjects	10	3786	378.60	—
F × *S*	90	671	7.46	—
SW × *S*	40	142	3.55	—
F × SW × *S*	360	1291	3.59	(error term)

Note.—Types C through G only.
* $p < .01$.

RESULTS

Table 1 gives the factorial analysis of variance and shows form and SW interacting significantly (.01 level of confidence) for the symbolic digits. The different forms did not give significantly different visibilities when stroke width was held constant. Form changes and SW variations together accounted for differences in perceptibility of the configurations. In other words, increased perceptibility of digits can be obtained by varying the SW emphasis of certain forms. Contrary to prior studies, the symbolic digits tended to be more perceptible.

DISCUSSION

The configuration of SW variations and form that was more perceptible varied greatly. No regular patterns appeared. In short, each digit configuration is a strictly unique event. The part of a digit requiring emphasis depends on that digit.

A few other clues can be pointed out. The factor "openness" suggested by other experi-

ments is the same for all symbolic types so its validity as a perception factor is questionable. The results demonstrate that varying SW within a digit may increase visibility. In most cases the symbolic configurations were more perceptible than either of the conventional types. Comparing some of the digits, the variation and lack of consistency from form to form with the best SW configurations is demonstrated. The differences in configuration in arabic and symbolic 7's are small; however, there were differences in the visibility. This raises a question on the use of anything other than straight lines for digit construction. The differences in narrow overall stroke and the wider overall stroke configurations give evidence that for some digit forms a narrower ratio is better and for others a wider stroke gives greater visibility. This would relate to the interaction effects demonstrated. Simplicity of form shows up as an asset to perception in several cases. For the 0, all the symbolic types were significantly more visible than the two conventional arabic 0's. The only difference in the shape of the conventional and symbolic 0's is that the former is circular and the latter is square. Compar this with the curved versus straight 7 results noted above.

Consider the problem of the interaction of stroke width and the shape of the digit. The data show that the same shaped digits had different levels of visibility depending upon the particular emphasis of SW made. The determination of interaction effects of these two variables shows they do significantly effect the legibility of the symbolic digits. Whereas in previous studies stroke width was varied the same overall each figure, and overall each set of 10 digits as well, we allowed for a variation within each set, and within each numeral also.

The results are at odds with previous studies. Previously stroke width (being the same overall each digit in other research designs) was secondary in effect to form when certain overall SW/H ratios were observed. The present studies show that varying the stroke width within each digit can increase visibility, and that form alone is not the determiner. Interaction of form and stroke width has a significant influence on the perceptibility also. The contributions of form, stroke width, and the interaction take no regular pattern.

The significant interaction of form and SW makes an interpretation of the effects of either difficult. Differences in form must have some influence on perceptibility however. The interaction of form and SW presumes some significance of form. Some types of a particular form were more significantly visible than other types. An inspection of the results seems to show that the stroke width effects have more influence on perception than the changes in form. This interaction brings many prior studies into a different light, since previously only one or the other (form or SW) was varied, not both. Established optimal SW/H ratios are possibly in need of a recheck.

Other studies typically determined the best stroke width to height ratios for an *entire set* of 10 digits, but in the present study these were varied from digit to digit. The results indicate, however, that some digits were increased in visibility by varying the boldness of the stroke width *within* individual digits, representing a major departure from past methodological and theoretical approaches to the study of numeral visibility. There is seemingly no reason, other than an esthetic one, to maintain a constant stroke width for any one digit or any set of digits. The results of the study show that *within variable stroke width does increase numeral visibility.*

Summary and Conclusions

As the complexity of visual displays and the demands of the machine upon the man increase, there is a real danger that the requirements will surpass the abilities of the human operators. One area of research in human factors concerns methods of presenting information to the operator in the most perceptible form. One small but important aspect in this problem is a design of digits to be used in the visual displays.

The results of previous research on digit design were summarized as follows: The optimum stroke width to height ratios have been apparently determined, angularity of digit shape increases visibility, and the openness of digits aids the discrimination.

Certain questions about digit design were approached in this study: What are the relative visibilities of conventional and symbolic digits? Do form and stroke width interact? And, most importantly, will varying the stroke

width within a set of digits and also within any particular digit increase the perceptibility?

The problem was to determine the effects of the above factors through a systematic variation of the stroke *within* a set of symbolic digits (straight-line, angular forms constructed from an eight-element matrix having cognates in the conventional digits).

The *S*-to-stimulus distance was varied by decreasing intervals until the *S* could correctly identify the digit. Eight types of digit configurations plus conventional arabic numerals were used. The 80 digits were presented, one at a time, in random order to the *S*s. Data were recorded in terms of the maximum distance at which the *S* could correctly identify the digit.

The results indicated that the symbolic numerals were significantly more visible than the conventional digits, contrary to prior reports. Form and stroke width showed a significant interaction, supporting the conclusion that stroke width could vary *within* a set of digits or even within a single digit so as to increase the discriminability.

We can generalize to say that each digit must be designed on an individual configuration basis taking form, stroke widths, and their combinations into account. Two factors must be considered: (*a*) absolute recognition of the digit as such, (*b*) differentiation (by relative emphasis) of features that suggest similarity (hence ambiguity) of one digit each of the others.

Questions still unanswered are: what are the effects of decreasing illumination to near thresholds, tachistoscopic presentation, and the relative perceptability of groups of the symbolic digits?

REFERENCES

Allusi, E. A., & Martin, H. B. An informational analysis of verbal and motor responses to symbolic and conventional arabic numerals. *J. appl. Psychol.*, 1958, **42**, 79–84.

Allusi, E. A., & Muller, P. F. Rate of information transfer with seven symbolic visual codes: Motor and verbal responses. *USAF WADC tech. Rep.*, 1956, No. 56-226.

Berger, C. Stroke-width, form and horizontal spacing of numerals as determinants of the threshold of recognition. Part I. *J. appl. Psychol.*, 1944, **28**, 208–231. (a)

Berger, C. Stroke-width, form and horizontal spacing of numerals as determinants of the threshold of recognition. Part II. *J. appl. Psychol.*, 1944, **28**, 336–346. (b)

Brown, F. R. A study of the requirements for letters, numbers, and markings to be used on trans-illuminated aircraft panels: IV. Legibility of uniform stroke capital letters as determined by size and height to width ratio and as compared to garamond bold. *USN Air Material Cent. Rep.*, 1953, No. TED NAM EL-609, Part 4.

Brown, F. R., & Lowery, E. A. A study of the requirements for letters, numbers and markings to be used on trans-illuminated aircraft control panels: I. The effect of stroke width upon the legibility of capital letters. *USN Air Material Cent. Rep.*, 1949, No. TED NAM EL-609, Part 1.

Brown, F. R., Lowery, E. A., & Willis, M. P. A study of the requirements for letters, numbers, and markings to be used on transilluminated aircraft control panels: III. The effect of stroke width and form upon the legibility of numerals. *USN Air Material Cent. Rep.*, 1951, No. TED NAM EL-609, Part 3.

Chapanis, A., Garner, W. R., & Morgan, C. T. *Applied experimental psychology*. New York: Wiley, 1949.

Cohen, J., & Webb, I. B. An experiment on coding of numerals for tape presentation. *USAF WADC tech. Rep.*, 1953, No. 54-86.

Craik, K. J. W. Instrument lighting for night use. *Air Ministry Flying Personnel Res. Committee Rep., Lond.*, 1941, No. 342.

Forbes, T. W., & Holmes, R. S. Legibility distances of highway destination signs in relation to letter height, letter width, and reflectorization. *Proc. Highway Res. Bd.*, 1939, **19**, 321–335.

Kuntz, J. E., & Sleight, R. B. Legibility of numerals: The optimal ratio of height to width of stroke. *Amer. J. Psychol.*, 1950, **63**, 567–575.

Lansdell, H. Effects of form on the legibility of numbers. *Canad. J. Psychol.*, 1954, **8**, 77–79.

Loucks, R. B. Legibility of aircraft instrument dials: A further investigation of the relative legibility of tachometer dials. *USAF Sch. Aviat. Med. proj. Rep.*, 1944, No. 1. (Project No. 286)

Mackworth, N. H. Legibility of raid block letters and numbers. *Air Ministry Flying Personnel Res. Committee Rep., Cambridge U.*, 1944, No. 423(s).

McLaughlin, S. C. Configuration and stroke-width in numeral legibility. Unpublished master's thesis, Tufts College, 1948.

Shapiro, H. B. Factors affecting the legibility of digits. *Amer. Psychologist*, 1951, **6**, 364. (Abstract)

Soar, R. S. Height-width proportion and stroke width in numeral visibility. *J. appl. Psychol.*, 1955, **39**, 43–46. (a)

Soar, R. S. Stroke width, illumination level, and figure-ground contrast in numeral visibility. *J. appl. Psychol.*, 1955, **39**, 429–432. (b)

Soar, R. S. Numeral form as a variable in numeral visibility. *J. appl. Psychol.*, 1958, **42**, 158–162.

(Early publication received June 16, 1961)

Journal of Applied Psychology
1961, Vol. 45, No. 6, 369–375

THE INTERACTION OF INFORMATION DISPLAYS WITH CONTROL SYSTEM DYNAMICS IN CONTINUOUS TRACKING[1]

R. W. OBERMAYER, W. F. SWARTZ, AND F. A. MUCKLER
Martin Company

Perhaps the most widely used display modes in human tracking studies have been following and compensatory information displays. In the compensatory display mode, the operator attempts to align a moving index with a fixed command index. In this case, the moving index is the sum of the operator's control movements, control system dynamics, and some arbitrary forcing function. In the following display technique, on the other hand, the forcing function is programed directly to the command index providing a moving profile of desired performance. The second moving index then becomes directly the resultant of the operator's control output. In both cases, of course, the objective is to null the error between the two display indices. Experimental comparisons of following and compensatory displays have been predominantly in favor of the following display (e.g., Poulton, 1950; Senders & Cruzen, 1952), presumably due to the additional information available to the operator.

However, since stimulus and response are clearly interdependent in a man-machine feedback control system, it is not surprising to find instances where the effects on tracking performance of type of information display and control system dynamics interact (e.g., Chernikoff & Taylor, 1957). Theoretically, one might expect that under certain dynamic conditions the advantages of the following display might disappear. If a control system includes integrators, the system output may continuously change without a corresponding continuous change of the operator's control force. Relieving the operator of the necessity of applying control force continuously, or in some time-sequenced pattern, has been termed "unburdening" by Birmingham and Taylor (1954). Based on the assumption that the human operator will perform best when his function is simplest, unburdening may result in a reduction of complexity in both display and control requirements. The compensatory display could be superior if the information contained in the instantaneous amplitude of the displayed error is sufficient to specify the response which reduces the error to zero. As Chernikoff and Taylor (1957) state: "The *S* would not have to take account of target velocities or error velocities nor would he have to predict in any fashion." Their data clearly suggest that the frequency of the course forcing function and variations of control system dynamics markedly affect the efficiency of compensatory and following tracking.

To provide additional clarification, this study was undertaken to investigate the interaction between modes of information display and control system dynamics with a low frequency simple sine wave. Both compensatory and pursuit modes of information display are compared where the operator is provided with direct control of successively higher derivatives of the control system output.

METHOD

Apparatus

Display. The display elements were a horizontal line and an inverted T presented on a 5-inch Dumont Dual Beam CRT. The horizontal line was the command index and was 2 inches long. The inverted T was the subject-controlled index. It consisted of a horizontal line 1 inch long with a vertical line, .5 inch long, bisecting the horizontal component. Both symbols were centered on the CRT and moved only in the vertical dimension.

When using the following display mode, the inverted T was controlled by the subject and the hori-

[1] This work was supported in part by the United States Air Force under Contract No. AF 33(616)-5472, monitored jointly by the Aerospace Medical Laboratory and the Flight Control Laboratory, Directorate of Advanced Systems Technology, Wright Air Development Division, Wright-Patterson Air Force Base, Ohio, as a part of the Air Force Control-Display Integration Program.

zontal line was positioned by a course generator. The compensatory mode consisted of the horizontal line being used as a reference on the center of the screen, while the inverted T was driven by the difference between the course generator output and the control system output.

Control. A center-positioned joy stick was used. It was 1 inch in diameter and 22 inches long. It was possible to move the top of the control stick in an arc of plus or minus 6 inches from the stick's null position by forward or rearward movements, respectively. The control stick was spring loaded so that it would return to within ¼ of an inch of its null position. A breakout force of .5 pound was required to displace the stick from its null position. The force-displacement gradient was linear and was 1 pound per inch. An analog deadband circuit was incorporated into the experimental apparatus to correspond to plus or minus ¼ of an inch of stick displacement.

A displacement of the control stick resulted in a proportional position, rate, or acceleration of the controlled index, i.e., control with no dynamics, one integrator, and two integrators, respectively. Two inches of stick movement resulted in 1 inch, 2 inches per second, or 8 inches per second squared of controlled index displacement for position, rate, and acceleration, respectively. All control system dynamics resulted in the same controlled index displacement at .5 second for a step function displacement of the control stick at the control gains chosen. A forward displacement of the control stick forced the controlled index downward and a rearward displacement forced the controlled index upward.

Course. A sine wave course, acting as a forcing function, was used in the experiment. The course had a frequency of 2-2/9 cpm, maximum amplitude of 2 inches, and a period of 27 seconds. The course was replicated with fixed amplitude, frequency, and direction from zero for each experimental trial.

Scoring. Measures of six dependent variables were obtained in the experimental situation:

1. The Average Absolute Error (AAE), or integrated error, was the average difference in inches between the command and controlled indices, without consideration of the sign of the differences, for each scoring period.

2. The Root Mean Square Error (RMS) was the square root of the integrated squared display error, in inches, which was obtained for each scoring period.

3. The Time-on-Target (TOT) was the time, in seconds, that the controlled index was within a tolerance band of plus or minus 1/10 of an inch of the command index.

4. Hits was the number of times that the error between the command index and the controlled index became less than plus or minus 1/10 of an inch.

5. Average Absolute Control Movement (AACM) described the average distance in inches along the arc of the control stick handle that the control stick was moved for each scoring period.

6. The Control-Stick Count estimated the number of times that the control stick was moved through the stick deadband of plus or minus ¼ of an inch.

Subjects

Nine adult males served as subjects in the experiment. They were selected from a broad population of professional pilots and engineers. In addition, the subjects were college graduates and varied between 25 and 40 years of age.

The treatment × treatment × subjects model of the analysis of variance was used (Lindquist, 1956, p. 237). There were six treatment combinations as a result of two levels of display modes and three levels of control dynamics. This design was used to increase the precision of the experiment by partitioning the variance due to intersubject differences. An interesting by-product of this design was the possible examination of the interaction of subject performance levels with each of the treatments.

Procedure

Each subject was placed in a completely enclosed mockup of a fighter-type cockpit facing the CRT display which was approximately 28 inches from the subject's eyes. A low intensity white light was placed inside the mockup to provide ambient illumination. The brightness of the light, fixed throughout the experiment, was sufficient to illuminate the panel.

An instruction and practice period was held prior to administration of the experimental treatments for each subject. The subjects were thoroughly briefed as to the nature of following and compensatory modes of information display and position, rate, and acceleration control system dynamics. The practice consisted of presenting each treatment combination of modes and dynamics to the subject for two trials. The sine wave course was used.

The experiment consisted of randomly presenting each of the six treatment combinations of modes and dynamics to each of the nine subjects. There were 5-minute rest periods between each treatment combination. Ten trials of each treatment combination were administered to each subject. Each trial was 1 minute in length with a 30-second rest period between trials. A yellow light on the panel was automatically lighted to warn the subject that 10 seconds remained before the beginning of the next trial. During the 30-second rest period, trial scores for each of the dependent variables were obtained at the experimenter's console. The overall time for the subject to complete the task was approximately 2 hours.

Results

Performance Scores

A total of 540 performance scores were obtained for each measure. Ten trials of each treatment combination were presented to the subject; however, only the last five trials of each combination were retained for examina-

tion in order that stable data might be obtained from each subject. Retaining only half the total observations resulted in each measure being based upon 270 scores.

Figure 1 presents the mean tracking performance data for the AAE and RMS measures; each mean was based upon 45 trials. The treatments are ranked along the abcissa of the figure in ascending order of treatment mean error magnitude from left to right. In Figure 1, a decrement of performance is noted for each change of control system dynamics to a higher derivative. Further, performance with following mode ranked better than compensatory mode for position control in both of the measures. However, this relation between following and compensatory was reversed for the rate and acceleration dynamics.

Figure 2 presents the data for the TOT and Hits measures. TOT decreased in value as the AAE and RMS increased in value. An examination of the Hits measure reveals that there were fewer tolerance band crossings for the rate control than for either the position or acceleration control. It becomes apparent, then, upon examining the TOT values, that the subjects tended to be on target for longer periods per time for the rate control than for position control even though better TOT performance was obtained for the position control. When compared to rate control, a relatively higher number of hits for the acceleration control

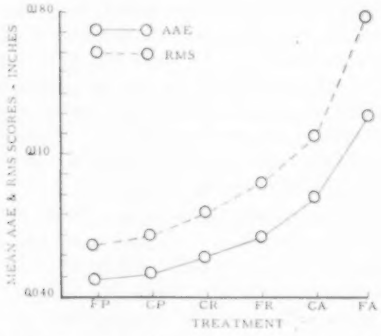


FIG. 1. Mean tracking performance data of each treatment for the AAE and RMS measures. (F = following display, C = compensatory display, P = position control, R = rate control, A = acceleration control.)

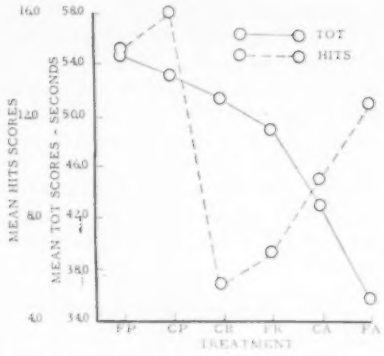


FIG. 2. Mean tracking performance data of each treatment for the TOT and Hits measures.

did not result in better TOT performance, however.

The Control-Stick Count measured the number of times that the control stick was moved through the null position; the results are shown in Figure 3. As the derivative of control is increased, Control-Stick Count decreases slightly, then increases considerably with acceleration control dynamics. It would appear from these data that the subjects response was not in the form of discrete positioning movements, for if this were the case Control-Stick Count would increase as the derivative of control dynamics was increased. Also, while it can be computed that the subjects could achieve zero error by employing a sinusoidal response, this too was not the case since otherwise no differences in Control-Stick Count would result. While no specific measure of the form of the subject's response was taken, it was observed that with acceleration dynamics the response of many subjects took the form of a high frequency, low amplitude signal (frequently termed "dither") superimposed on a low frequency, high amplitude signal.

The AACM, as shown in Figure 3, decreased as a direct function of increasing derivative of control system dynamics. Considerably less control stick movement resulted with rate and acceleration dynamics than with position control dynamics. There was very little difference in AACM for the two modes of information display.

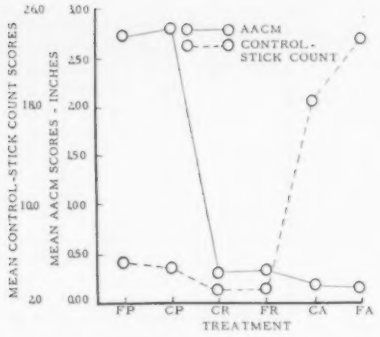


FIG. 3. Mean tracking performance data of each treatment for the AACM and Control-Stick Count measures.

Statistical Analysis

Analysis of Variance Results. The measures of AAE, RMS, and TOT were statistically examined by means of the treatment × treatment × subjects model of the analysis of variance (Lindquist, 1953). Prior to analysis, the data for each of the three measures had to be transformed to satisfy the assumptions of the analysis of variance. The appropriate transform for each measure was determined on the basis of the normality plot for the respective distribution. It was determined that the AAE and RMS measures should be transformed logarithmically (log AAE and log RMS) and that the TOT measure should receive a tangent transformation tan 1.5 (TOT-10). Upon completion of the transformations, homogeneity of variance for each measure was examined (Odeh & Olds, 1959). For each $\bar{X}$ used in the analysis of variance there is associated an s^2. Using the procedure established by Odeh and Olds, an identical analysis of variance was conducted on the quantity log s^2. By using this method, the distribution of each source of variance was analyzed separately with respect to heterogeneity. The results of this analysis are presented in Table 1 in conjunction with the summary of the analysis of variance results.

TABLE 1

SUMMARY OF THE ANALYSIS OF VARIANCE RESULTS FOR THE MEASURES OF AAE, RMS, AND TOT

Source	Significance level: Measure AAE	RMS	TOT
Modes	5%*	*ns*	*ns*
Dynamics	1%	1%	1%**
Subjects	1%	1%	1%*
M × D	1%	5%	1%
M × S	*ns*	*ns*	5%
D × S	1%	1%	1%

* Heterogeneous at 5% level as determined by the method of Odeh and Olds (1959).
** Heterogeneous at 1% level as determined by the method of Odeh and Olds (1959).

TABLE 2

RESULTS OF THE MULTIPLE RANGE TEST FOR THE M × D INTERACTION OF THE AAE, RMS, AND TOT MEASURES

Measure	Range comparisons					
AAE	FP	CP	CR	FR	CA	FA
RMS	FP	CP	CR	FR	CA	FA
TOT	FP	CP	CR	FR	CA	FA

Note.—The underline denotes no significant difference at the 5% level.

In Table 1, the most interesting results are displayed in the Modes × Dynamics interaction. It is noted that the significant interaction, whose distribution was homogeneous at the 5% level, has been obtained in each of three measures. A significant Dynamics × Subjects interaction was found, also, for all measures, but the Modes × Subjects interaction was significant for only the TOT measure. Unfortunately, it cannot be determined whether the relation between display modes and control dynamics was also dependent upon subject characteristics since the M × D × S interaction was nontestable.

Multiple Group Comparisons. Table 2 presents the results of Duncan's new multiple range test (Duncan, 1955) of the M × D interaction in summary form for each of the measures. All multiple comparisons of the treatment means were conducted at the 5% level. An underline denotes no significant difference between those treatment means at the 5% level; all other comparisons are significant. In addition, the treatments are ranked,

by the magnitude of their means, in ascending order of error from left to right. It is noted that there was no difference in the statistical outcomes in the RMS and TOT measures and very little between the AAE and RMS-TOT measures. The treatment ranking was the same for all three measures.

Examining the following and compensatory modes for common conditions of position, rate, and acceleration dynamics in Table 2 gives statistical bearing to the data contained in Figure 1. Compensatory was nonsignificantly different from following mode for position and rate control, but was significantly superior to following for acceleration control in each of the three measures.

Examining position, rate, and acceleration dynamics for common conditions of following and compensatory modes discloses that position, rate, and acceleration were each significantly different from one another for the following mode with position control being the best and acceleration the worst; however, only the acceleration control was significantly different from the position control for the compensatory mode. It is noted that these results, too, were the same for all three measures.

Individual Subject Comparisons. Table 3 shows the relationship of position, rate, and acceleration dynamics for each subject within the significant D × S interaction. Mean performance scores, when averaged over all subjects, indicated that a decrement in performance occurred with changes from position to rate to acceleration control for each measure. This is demonstrated in Figure 1. However, Table 3 demonstrates that performance for each individual subject was not affected, in all cases, in the manner which is indicated by averaging over all subjects. The same results were not obtained for each of the measures.

TABLE 3

COMPARISON OF POSITION, RATE, AND ACCELERATION DYNAMICS FOR EACH SUBJECT

	Range comparison		
	Measure		
	AAE	RMS	TOT
1	R P A	P R A	P R A
2	R P A	P R A	R P A
3	P R A	P R A	P R A
4	P A R	P A R	P A R
5	P R A	R P A	R P A
6	P R A	P R A	P R A
7	R P A	R P A	R P A
8	P R A	P R A	P R A
9	P R A	P R A	P R A

Note.—An underline denotes no significant difference at the 5% level.

DISCUSSION

It is interesting in light of the abundance of research proclaiming the superiority of the following mode of information display by a factor of 2 to 1 or more, that in no condition of this experiment was the following display found to be superior to the compensatory display. Both minimum and maximum tracking error in this study occurred with the following display. This is taken as further evidence that with low frequency courses the subject needs little more than simple error information provided by the compensatory display. Indeed, it is perhaps surprising that the results of this study were not more favorable to the compensatory display, since it is suggested by Birmingham and Taylor (1954) that in cases where a simple stimulus-response relationship exists, the additional information provided by the following display may contribute to degraded system performance.

Using a complex course at about the same frequency as the simple sine wave in the current study (2-2/3 + 4-4/9 + 6-2/3 cpm), Chernikoff and Taylor (1957) found rate control to be superior to position control with both modes of information display. In the current study no reliable difference between position and rate dynamics was found with the compensatory display, and rate control was significantly inferior with the following display. One would expect with a regular and highly predictable sine wave course that the benefits of unburdening would be more prominent than with a complex course. A possibly critical difference between the two studies was that the average course velocity in the current study was higher (Bowen & Chernikoff, 1958). It would appear that there is reason to suspect that course variables exist, in ad-

dition to course frequency, which interact with modes of information display and control system dynamics.

It can be computed that if the output of each subject were sinusoidal and of the same frequency as the course, zero tracking error could result for the course used in this study. Unfortunately, there were no data collected which would directly specify the form of the subject's response; however, it can be inferred that the subject's response was not totally in the form of simple corrective movements and since tracking error was quite small, it may be assumed that a sinusoidal response was closely approximated. While the form of the desired subject response is sinusoidal for each order of derivative control, the phase relationship between the desired control movement and the sinusoidal course movement differ with control system dynamics. For acceleration control dynamics, the desired subject's control movements are of the same sinusoidal form but in exactly the opposite direction from that which would achieve zero error with position control dynamics (Ely, Bowen, & Orlansky, 1957, p. 15). The "natural" control sensing was used for position control dynamics; therefore, the control sensing for acceleration control dynamics may be termed the "unnatural" sensing. Green, Norris, and Spragg (1955) suggest that due to the fact in compensatory tracking the resulting movement of the target is the result both of the subject's input and the course generating system, the subject does not get a clear feedback of the results of his own inputs, and therefore, these responses may be less subject to influence by changing the display-control relationships. The following display does provide separate display of the results of the subject's responses; presumably, the subject would be well informed of the effect of his responses, and, in addition, could compare his stick motion to the motion of the course symbol. It is, therefore, possible that the superiority of the compensatory display with acceleration control dynamics may be in part due to the existence of an unnatural control-display relationship with these dynamics.

Since there are few identifying characteristics of the subjects of this experiment, it is futile to attempt to correlate the different trends in performance found among individuals to specific subject characteristics. However, that the results of tracking studies may be dependent on the particular individuals employed is not a new finding. Walston and Warren (1954) in attempting to measure a mathematical transfer function for each subject found that quite different tracking techniques were used by some individuals. Muckler (1960), for example, in investigating tracking performance with oscillatory transients, showed that trends in the performance data depended on the individual subjects and that even reversals in the overall trend may result. He concluded that the performance scores obtained were quite dependent on the control techniques used by the individual subjects and that the task techniques must be considered in the interpretation of the data. In the present study, three different rankings of dynamics were found among nine subjects, constituting six different statistical outcomes. It is clear, then, that the data of this study might have been changed considerably by judicious sampling of the nine subjects.

A control system gain was arbitrarily chosen for each condition of control system dynamics used in this study. While 2 inches of control movement produced display changes of 1 inch, 2 inches per second, or 8 inches per second squared for position, rate, and acceleration dynamics, respectively, a wide range of these parameters might have been equally acceptable for the purposes of this study. The conditions of dynamics used, then, are best considered as control gain-dynamics combinations. There is no guarantee that the control gain-dynamics used were optimum; indeed, the optimum is likely a function of display mode. It is entirely possible, therefore, that the results of this experiment are not representative of all control gain-dynamics combinations. It is suspected that in part the considerable reduction in AACM scores represents an attempt by the subjects to achieve system stability by reducing his overall gain; however, since the amount of control movement is also dependent upon the control gain associated with each control dynamic, little of a concrete nature can be stated here. The unavoidable conclusion is that only additional

research can resolve the influence of control gain in this study.

Summary

Nine skilled subjects performed single dimension tracking with all combinations of following and compensatory modes of information display and position, rate, and acceleration control system dynamics. The forcing function was a slow frequency, simple sine wave course. Tracking performance was found to be dependent upon the interaction between display modes and control system dynamics. Compensatory tracking was equivalent to following tracking for position and rate control, but was significantly superior to following tracking for acceleration control. For the compensatory mode, position and rate control tracking performance was statistically equivalent; however, both position and rate control were significantly superior to acceleration control. For the following mode, all dynamics conditions were significantly different with position control demonstrating highest and acceleration control lowest performance levels. Examination of significant subject interactions with treatments showed deviation in individual cases from the averaged group data.

REFERENCES

Birmingham, H. P., & Taylor, F. V. A human engineering approach to the design of man-operated continuous control systems. *USN Res. Lab. Rep.*, 1954, No. 4333.

Bowen, J. H., & Chernikoff, R. The effects of magnification and average course velocity on compensatory tracking. *USN Res. Lab. Rep.*, 1958, No. 5186.

Chernikoff, R., & Taylor, F. V. Effects of course frequency and aided time constant on pursuit and compensatory tracking. *J. exp. Psychol.*, 1957, **53**, 285–292.

Duncan, D. B. Multiple range and multiple *F* tests. *Biometrics*, 1955, **11**, 1–42.

Ely, J. H., Bowen, H. M., & Orlansky, J. Man-machine dynamics. *USAF WADC tech. Rep.*, 1957, No. TR 57-582.

Green, R. F., Norris, E. B., & Spragg, Shirley D. Compensatory tracking performance (modified SAM two-hand pursuit test) as a function of the directions and planes of movement of the control cranks relative to the movement of the target. *J. Psychol.*, 1955, **40**, 411–420.

Lindquist, E. F. *Design and analysis of experiments in psychology and education*. Boston: Houghton Mifflin, 1953.

Muckler, F. A. Man-machine tracking performance with short period oscillatory control system transients. *USAF WADD tech. Rep.*, 1960, No. TR 60-3.

Odeh, R. E., & Olds, E. G. Notes on the analysis of variance of logarithms of variances. *USAF WADC tech. Note*, 1959, No. TN 59-82.

Poulton, E. C. Two pointer and one pointer displays in tracking. *Med. Res. Council Rep.*, 1950, No. APRU-150/50.

Senders, J. W., & Cruzen, Marianne. Tracking performance in compensatory and pursuit tasks. *USAF WADC tech. Rep.*, 1952, No. TR 52-39.

Walston, C. E., & Warren, E. C. A mathematical analysis of the human operator in a closed-loop control system. *USAF Personnel Train. Res. Cent. tech. Rep.*, 1954, No. TR 54-96.

(Early publication received July 20, 1961)

Journal of Applied Psychology
1961, Vol. 45, No. 6, 376–380

CROSS-VALIDITY OF PROCEDURES FOR SELECTING LIFE INSURANCE SALESMEN[1]

PETER F. MERENDA,[2] WALTER V. CLARKE, AND CHARLES E. HALL[3]

Walter V. Clarke Associates

An extensive study of personality and biographical factors as predictors of success-failure of life insurance agents has shown that both personality characteristics (Merenda & Clarke, 1959a) and personal history variables (Merenda & Clarke, 1959b) are related to success in selling ordinary life insurance. The previous findings have further demonstrated that the predictive efficiency of selecting life insurance salesmen can be enhanced by the combined use of both sets of predictor variables.

An earlier study of Clarke (1956b) using the same personality assessment instrument employed in this study revealed evidence of discriminatory power of the personality variables in distinguishing between high and low producers among life insurance agents beyond 2 years after hire. Another parallel study by Wallace, Clarke, and Dry (1956) failed to show the same significant discriminations between first and second year successful and unsuccessful agents.

The study presented in this paper reports on the cross-validity of the prediction system used in the earlier validity study (Merenda & Clarke, 1959b) administered to an independent sample of life insurance agents. The main intent of the study was to determine whether the power of prediction found in the validity study (Merenda & Clarke, 1959b) would hold up on cross-validation. A further purpose was to obtain data which would help clarify the apparent discrepancies found in the earlier studies of Clarke (1956b) and Wallace, Clarke, and Dry (1956).

[1] The major portions of this paper were presented at the Annual Meetings of the Midwestern Psychological Association, St. Louis, Missouri, April 29, 1960.

[2] Now at the University of Rhode Island.

[3] Now at the Rhode Island State Department of Education.

METHOD

Subjects. Subjects in this cross-validation study were 535 financed male life insurance agents hired by an ordinary life insurance company between January 1, 1955 and July 31, 1956. The earlier validation sample was comprised of 522 agents hired by this same company in the period between September 1, 1950 and December 31, 1954.

Predictors. Three sets of predictors of success as life insurance salesmen were employed with the validation sample. The first was a self-concept personality assessment inventory. The particular instrument was the Activity Vector Analysis (Clarke, 1956a) or AVA, as it is most commonly called. The AVA was administered at the time of application for the position of salesman. The second set was composed of five personal-social history items contained in the application blank completed by the salesmen prior to hire. A third predictor of age of the agent at the time of application for the position of life insurance salesman was used in the earlier validation study.

Criteria. Two sets of criterion standards were used. The first criterion was a Success-Failure dichotomy determined by combining dollar production with job status within the company at the end of a 3-year period after hire.

A successful agent ($N = 97$) is one who:

1. Meets his Training Allowance Program quotas or achieves $200,000 production in his first year and at least $300,000 in either his second or third year; or
2. Is advanced to a supervisory or management position within the company; or
3. Leaves the company to become an agent, supervisor, or general agent of another company before the end of the third year if he achieves the production goals outlined above.

An unsuccessful agent ($N = 438$) is one who:

1. Fails to reach the production goals outline, whether or not he remains as an agent with the company; or
2. Has had his contract terminated by the company; or
3. Leaves the insurance industry within a 3-year period.

The second criterion was the new business volume (face value of insurance policies) at the end of the first, second, and third years.

Procedure. In the validity study, discriminant analysis was applied to the problem of providing maximum separation between the 522 agents di-

TABLE 1

RELATIVE PREDICTIVE EFFICIENCIES OF AVA, PERSONAL HISTORY, AND AVA PLUS PH FOR VALIDITY AND CROSS-VALIDITY SAMPLES

Sample	*N*	Number of agents failing to meet standards						Percentage rejected by these standards	
		AVA alone		PH alone		AVA + PH		3-year	3-year
		S	U	S	U	S	U	S	U
Validity	522	2	39	17	134	19	156	17	38
Cross-validity	535	6	60	4	73	10	121	10	28

chotomized on the first criterion into 108 successful and 414 unsuccessful salesmen. Separate analyses were made for the personality, personal history, and age predictors. The discriminant functions for each predictor set proved to be statistically significant. With respect to the age predictor, it was found in the validity study that applicants who were younger than 25 or more than 45 years old proved to be substantially less successful than those who were within this age group.

The discriminant weights derived on the validity sample were applied to the 535 subjects in the cross-validity sample. Applying the first criterion standard to the sample of 535, the group was dichotomized into 97 successful and 438 unsuccessful agents. Employing the same cut-offs indicated in the validation study as being the optimal discriminant scores for rejecting the greatest proportion of potentially unsuccessful agents at the sacrifice of a minimum of potentially successful salesmen, predictions were made as to the success-failure of subjects in the cross-validity sample. Also analyses were made of the relationship existing between the range of discriminant scores on the predictor sets and agent production studied separately for the personality and personal history variates. These cross-validities were established against the single criterion of paid production independently of dichotomization on the success-failure criterion. The discrimination results on the age predictor, however, were not upheld by the data of the cross-validity sample. Hence this variate was removed from the predictor set in the cross-validation procedure reported in this paper.

RESULTS AND DISCUSSION

The linear discriminant scores representing the AVA profiles yielded a mean of 12.37 and *SD* of 8.43 for the successful group, and a mean of 9.08 and *SD* of 8.55 for the unsuccessful group. For the total cross-validity sample, the AVA discriminant scores ranged from −10 to +33. The linear composite score distributions of the personal history variates yielded a mean of 108.64 and *SD* of 10.56 for the successful group, and a mean of 100.03 and *SD* of 9.80 for the unsuccessful group. For the total sample, the personal history discriminant scores ranged from 70 to 134.

Table 1 presents the results obtained in the cross-validation of the predictive efficiency of the predictor sets, employing multiple cut-offs. Any subject was considered to be a potential failure who failed to achieve at least a discriminant score of zero on the AVA set and a score of 91 on the personal history set of predictors.

It is noted in Table 1 that the validity of prediction of failure of a significantly great proportion of life insurance salesmen who were actually unsuccessful at the end of 3 years in the validity sample is upheld by the data of the cross-validity sample. While some shrinkage is noted in the percentage of unsuccessful agents who failed to meet the standards, a gain arose in the power of prediction of successful agents. A drop of 7 percentage points (from 17% to 10%) was noted in the proportion of successful agents who failed to meet the standards.

It is further noted in Table 1 that the ratio of unsuccessful salesmen to successful agents was greater for the cross-validity sample than in the validity sample. In the earlier sample, 108 of the 522 agents were considered to be successful at the end of 3 years, whereas in the later sample only 97 of 535 were successful in terms of the criteria of the study. Two other significantly different features were noted: the cross-validity sample contained more agents than the former sample who

TABLE 2

FIRST YEAR PRODUCTION YIELD OF PRODUCTIVE AGENTS BY CATEGORIES OF THE AVA DISCRIMINANT SCORE

Score range[a]	*N*	Median[b]	Mean[b]	σ
250– 349	26	250.00	250.00	177.75
150– 249	102	156.25	201.95	173.75
50– 149	182	172.22	203.00	160.40
−49– 49	96	130.77	160.95	135.15
−149– −50	19	116.67	127.65	105.70

Note.—The data of this table are based on only those agents who were still employed at the end of the first year.
[a] The original discriminant scores were multiplied by the constant 10 in order to eliminate the use of decimals in the values indicated in the score range.
[b] Paid production in thousands of dollars.

failed to pass the personality criterion, and fewer who failed to pass the personal history criterion.

Although fewer of the agents failed to pass the selection screen than was the case with the validity sample, a total of 121 unsuccessful agents in the cross-validity sample would have been rejected at the sacrifice of only 10 who proved to be 3-year successes. Over the 3-year period these 121 unsuccessful agents sold a yearly average of $40,730 life insurance (face value). The 10 successful agents sold a yearly average of $340,000. The costs to the company in the initial recruitment, training, and financial support of these financed agents constituted a far greater investment than the 3-year return from this group of 131 salesmen who would not have been hired had the selection system been in effect at the time of hire.

A further criterion in the form of paid production (face value of life insurance sold) was applied in the study.

Personality Variables

Tables 2–4 present the median and mean productions of agents who were actively engaged in selling insurance at the end of the first, second, and third years. In these tables, the sample subjects are classified according to AVA discriminant score categories regardless of classification on the success-failure criteria of the study.

The data of Tables 2, 3, and 4 reveal that there is a high relationship between scores obtained on the personality profiles and agent production at the end of each of 3 years. Both the median and mean production yields of agents who were actively engaged in selling life insurance at the end of 1, 2, and 3 years, respectively, appear to vary directly with the category of discriminant score ranges on the AVA. It is further noted in these tables that those agents with discriminant scores of 50 or above on the AVA sold, on the average, an amount of insurance which either closely approached or exceeded the minimum acceptable standards of production set by the company agency department. These standards were $200,000 at the end of the first year, and $300,000 at the end of either the second or third year. The data further show that those

TABLE 3

SECOND YEAR PRODUCTION YIELD OF PRODUCTIVE AGENTS BY CATEGORIES OF THE AVA DISCRIMINANT SCORE

Score range[a]	*N*	Median[b]	Mean[b]	σ
250– 349	16	300.00	384.40	326.05
150– 249	50	288.89	333.00	232.45
50– 149	90	288.89	287.20	212.35
−49– 49	40	200.00	263.75	217.80
−149– −50	5	200.00	174.50	134.15

Note.—The data of this table are based on only those agents who were still employed at the end of the second year.
[a] The original discriminant scores were multiplied by the constant 10 in order to eliminate the use of decimals in the values indicated in the score range.
[b] Paid production in thousands of dollars.

TABLE 4

THIRD YEAR PRODUCTION YIELD OF PRODUCTIVE AGENTS BY CATEGORIES OF THE AVA DISCRIMINANT SCORE

Score range[a]	*N*	Median[b]	Mean[b]	σ
250– 349	8	416.67	425.00	200.00
150– 249	35	292.86	350.70	217.25
50– 149	59	291.67	336.00	337.00
−49– 49	27	225.00	243.50	167.90
−149– −50	3	225.00	208.35	62.30

Note.—The data of this table are based on only those agents who were still employed at the end of the third year.
[a] The original discriminant scores were multiplied by the constant 10 in order to eliminate the use of decimals in the values indicated in the score range.
[b] Paid production in thousands of dollars.

agents with AVA profiles which were closest to those possessed by the successful category of agents in the validity study sold, on the average, substantially more life insurance than those whose personality profiles (as revealed by relatively low or negative discriminant scores) were incompatible with the successful agents in the validity sample.

Personal History Variables

The cross-validity of the personal history predictor set is demonstrated by the data of Tables 5, 6, and 7. Again, as with the cross-validity findings on the personality factors, a near perfect relationship was found to exist between the median/mean productions and category classification with respect to discriminant score range on the personal history set of variates. Those with the lowest discriminant scores tended to sell, on the average, considerably less life insurance than those with high scores. It is further noted from the data of these tables that the cut-off score of 91 which was indicated by the results of the earlier study to be an optimal criterion seems to be confirmed by the cross-validity findings. Those agents with scores below this cut-off sold, on the average, less than the criterion amount of life insurance in each of the 3 years. Also, by the end of the third year only 7 of the original 52 agents in this category were still employed by the company.

TABLE 5

First Year Production Yield of Productive Agents by Categories of the Personal History Discriminant Score

Score range	*N*	Median[a]	Mean[a]	σ
120 and above	21	208.34	232.15	217.28
110–119	58	228.57	258.62	184.86
100–109	153	162.50	193.30	152.41
90– 99	141	143.33	178.90	150.76
89 and below	52	111.11	138.46	112.29

[a] Paid production in thousands of dollars.

TABLE 6

Second Year Production Yield of Productive Agents by Categories of the Personal History Discriminant Score

Score range	*N*	Median[a]	Mean[a]	σ
120 and above	12	525.00	533.34	360.46
110–119	39	299.00	323.72	222.89
100–109	70	283.34	301.43	224.52
90–99	61	272.73	272.54	202.33
89 and below	19	99.00	180.27	173.88

[a] Paid production in thousands of dollars.

Summary and Conclusions

An independent sample of 535 ordinary life insurance agents was used to cross-validate a selection system originally validated on an earlier sample of 522 agents from the same company. The predictor sets were: (*a*) a personality profile measured by the Activity Vector Analysis and (*b*) a set of five personal history variables obtained in a weighted application blank. The criteria were: (*a*) 3-year agent success-failure determined on a multiple criterion basis incorporating agent production and company status and (*b*) dollar amount of face value of life insurance sold. The predictions were made on the basis of multiple cut-offs, i.e., any agent failing to pass either of the critical scores on the two predictor sets was identified as a potential failure. These cut-off scores were determined from the discriminant analysis results from the validity sample of 522 agents.

The data of this cross-validity study confirm the findings obtained in the original validation study. These are:

1. The AVA is a valid predictor of success-failure among life insurance agents.

2. Certain personal history measures are valid predictors among life insurance agents.

TABLE 7

Third Year Production Yield of Productive Agents by Categories of the Personal History Discriminant Score

Score range	*N*	Median[a]	Mean[a]	σ
120 and above	12	383.33	604.17	542.55
110–119	28	275.00	301.79	177.53
100–109	48	275.00	290.63	219.41
90– 99	37	285.71	307.43	218.79
89 and below	7	175.00	239.29	172.62

[a] Paid production in thousands of dollars.

The predictive validity of one of the measures (age) was not, however, upheld.

3. Combining AVA and personal history data enhances the predictive efficiency of these measures in determining success or failure of life insurance agents over a sustained period of time.

REFERENCES

CLARKE, W. V. The construction of an industrial selection personality test. *J. Psychol.*, 1956, **41**, 379–394. (a)

CLARKE, W. V. The personality profiles of life insurance agents. *J. Psychol.*, 1956, **42**, 295–302. (b)

MERENDA, P. F., & CLARKE, W. V. Activity Vector Analysis validity for life insurance salesmen. *Engng. industr. Psychol.*, 1959, **1**, 1–11. (a)

MERENDA, P. F., & CLARKE, W. V. The predictive efficiency of temperament characteristics and personal history variables in determining success of life insurance agents. *J. appl. Psychol.*, 1959, **43**, 360–365. (b)

WALLACE, S. R., CLARKE, W. V., & DRY, R. J. The Activity Vector Analysis as a selector of life insurance salesmen. *Personnel Psychol.*, 1956, **9**, 337–345.

(Received January 1, 1960)

Journal of Applied Psychology
1961, Vol. 45, No. 6, 381–387

EXPERIMENTAL EVALUATION OF BINARY CODES FOR CONSOLE DISPLAY

FRANK J. MINOR AND STANLEY L. REVESMAN

International Business Machines Corporation, Endicott, New York

This study experimentally evaluated three binary code schemas in terms of operator coding performance. The purpose of the evaluation was to aid in the selection of one of three alternative code schemas to be used on a data processing system console display. Information to be displayed in coded form would be a computer memory address displayed on an "address register display," and the contents of that address displayed on a "character register display." This information is required by the system programer to check the accuracy of a series of data processing instructions which he has assembled to serve as a data processing program. In installations of a given type such checking, or debugging, is accomplished by executing via the system console one program instruction at a time in its proper sequence. During this cycling process the programer makes careful note of the program instruction, the raw data, the data results and their corresponding memory addresses as displayed in coded form at the appropriate console character and address registers. This information is checked by the programer to locate errors of programing which if left uncorrected would result in invalid data processing.

The planned data flow logic of the system was amenable only to the three codes evaluated in this study, and would not permit modifications of the codes. The code schemas were:

1. Biquinary Code
2. Binary Coded Decimal
3. Combined Code

Codes 1 and 2 are currently in use for display purposes on data processing systems. Code 3 is a workable combination of the first two codes, but is not currently known to be in use for display purposes. Because of the market for which this system was designed, transfer of training was not an issue in this study. It is not expected that programers trained for existing stored program data processing systems would also perform programing for this new system.

Of the three code schemas, the Binary Coded Decimal was the most efficient in terms of cost and electronic simplicity for this specific system. The difference in efficiency however was not sufficient to justify eliminating the other two schemas from consideration. This study was conducted to determine what performance differences could be expected in terms of programer's performance at the console as a function of these code variations.

METHOD

Dependent Variables

The criterion for the codes evaluation was the programer's efficiency measured in terms of speed and accuracy of coding console address and character registers. The speed with which a programer can perform the coding task is critical since the data processing system is nonproductive during the debugging procedures at the console. When the hourly rental cost of such a high speed data processing system is considered, any reduction in system downtime makes for more efficient utilization of the system. Frequently there is more than one programer assigned to a system. Each programer writes and debugs his own programs.

The program debugging task at the console is not a primary task of any one programer, and is of an intermittent nature. However, total system downtime due to debugging by all programers assigned to the system could significantly reduce potential system computing time. The frequency and time spent by any one programer interrogating a system for program debugging purposes will vary as a function of the number and complexity of new programs he is required to write, his ability, and his preferred work methods. Barring very short and simple programs, newly written programs contain errors which require detection and correction.

Experimental Conditions

Each of the three possible code schemas facilitated the coding of 48 characters. A character could be displayed on the computer console by the illumination of one or more of a configuration of neon bulbs in the display register. For the experimental conditions

TABLE 1

BIQUINARY CODE SCHEMA

	Bit codes																				
	12	11	0	N	0	1	0 1	2 1	3	4 5	6 7	8 9
A	X					X	X		X			
B	X				X			X				
C	X					X		X				
D	X				X					X		
E	X					X				X		
F	X				X						X	
G	X					X					X	
H	X				X							X
I	X					X						X
&	X				X			X				X
-	X									X		X
□	X											
J		X					X	X				
K		X				X			X			
L		X					X		X			
M		X				X				X		
N		X					X			X		
O		X				X					X	
P		X					X				X	
Q		X				X						X
R		X					X					X
¢		X				X		X	X			
$		X							X		X	
*		X								X		X
/			X				X	X				
S			X			X			X			
T			X				X		X			
U			X			X				X		
V			X				X			X		
W			X			X					X	
X			X				X				X	
Y			X			X						X
Z			X					X				X
+			X			X			X			
,			X						X		X	
%			X							X		X
1				X			X	X				
2				X		X			X			
3				X			X		X			
4				X		X				X		
5				X			X			X		
6				X		X					X	
7				X			X				X	
8				X		X						X
9				X			X					X
0				X		X		X				
#				X			X		X			
@				X		X				X		

TABLE 2

BINARY CODED DECIMAL SCHEMA

	Bit codes								
	C	B	A	8	4	2	1
A		X	X				X
B		X	X			X	
C	X	X	X			X	X
D		X	X		X		
E	X	X	X		X		X
F		X	X		X	X	
G		X	X		X	X	X
H		X	X	X			
I	X	X	X	X			X
&	X	X	X	X		X	
-		X	X	X		X	X
□	X	X	X	X	X		
J	X	X					X
K	X	X				X	
L		X	X			X	X
M	X	X			X		
N		X	X		X		X
O		X	X		X	X	
P	X	X	X		X	X	X
Q	X	X		X			
R		X	X	X			X
¢		X	X	X		X	
$	X	X		X	X		
*		X	X	X	X		
/	X		X				X
S	X		X			X	
T			X			X	X
U	X		X		X		
V			X		X		X
W			X		X	X	
X	X		X		X	X	X
Y	X		X	X			
Z			X	X			X
+			X	X		X	
,	X		X	X		X	X
%			X	X	X		
1							X
2						X	
3	X					X	X
4					X		
5	X				X		X
6	X				X	X	
7					X	X	X
8				X			
9	X			X			X
0				X		X	
#	X			X		X	X
@	X			X	X		

the possible configurations of illuminated neon bulbs were simulated with paper and pencil coding tasks. The configuration of lights for each code structure was simulated by printed circles on the test sheets. A description of each code condition is as follows:

Code Condition 1 is an 11-bit (11 lights) Biquinary Code system of which 3 of the 11 lights are impulsed to denote a character. The code system as shown in Table 1 would be the basis for the character register display; a modification of this code system would be the basis for the address register display.

Code Condition 2 is a 7-bit Binary Coded Decimal system of which a variable number of neon bulbs averaging 3.33 per character are impulsed to denote a character. The code system as shown in Table 2 would be the basis for the character register display; a modification of this code system would be the basis for the address register display.

Code Condition 3 is a combination of Codes 1 and 2 such that the character coding schema for Code 1 would be the basis for the character register display; the address register coding schema for Code 2 would be the basis for the address register display (see Tables 1 and 2).

Design

A 3×3 experimental design with repeated measurements of independent groups was utilized. The three code schemas were analyzed under a series of three separate trials.

Subjects

Fifty-five naive subjects (33 males, 22 females) were employed in the study. None of the subjects had had previous experience with nondecimal number systems or codes such as those of the experiment. Subjects were paid volunteers from a junior technical college (Broome Technical Community College). All subjects were drawn from a population of students majoring in Business Technology. This particular population was selected to increase the homogeneity of subjects on task relevant variables (primarily clerical skills) and also to have subjects who were representative of the expected data processing system programers.

Two of the independent groups consisted of 18 subjects each; the other group, 19 subjects. Scheduling limitations precluded a completely random assignment of males and females across groups. Consequently, there was a disproportionate assignment of males and females to each group. A post facto analysis, however, supported the assumption of no sex differences for the coding task. To test for equality in general ability level between the groups, t tests were conducted between groups based on the cumulative scholastic average of each subject. The results of the t tests indicated no differences between groups.

Apparatus

Printed master code sheets which presented all letters of the alphabet, the numbers 0–9, and 12 special characters were made up for each code condition. The code sheets were used during training and testing of the subjects.

Paper and pencil tests were utilized in the experiment. These were designed to simulate coding problems which would occur on the data processing system console registers (see Figures 1 and 2 for examples). Three test booklets were used under each ex-

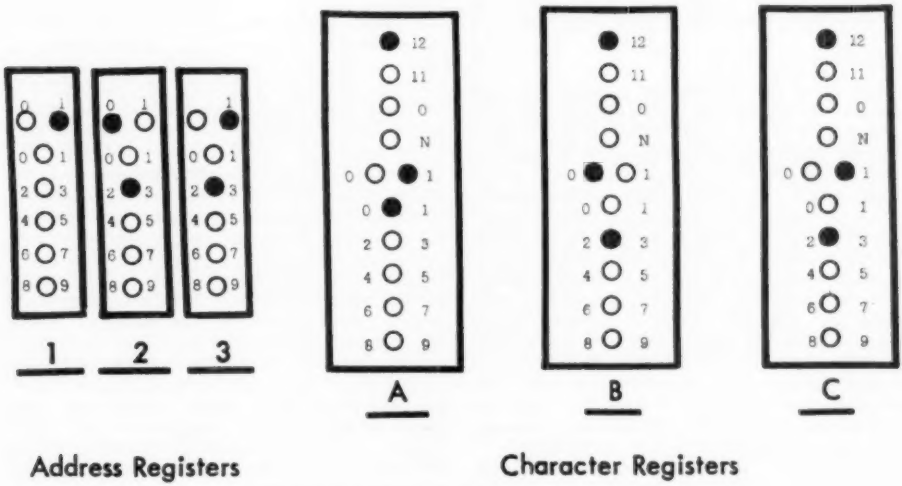


FIG. 1. Examples of simulated display registers for Biquinary Code.

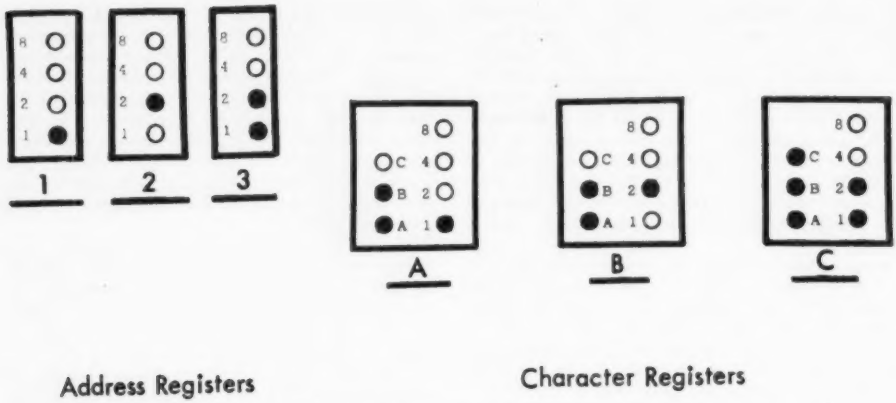


FIG. 2. Examples of simulated display registers for Binary Coded Decimal Code.

perimental code condition. Each test booklet consisted of 36 problems. Each problem is defined as consisting of a character register and an address register.

In one-half of the problems the subjects were required to encode by placing an X in the appropriate circles. In the remaining one-half of the problems the test items were presented with Xs already in the printed circles and the subjects were required to decode by placing the appropriate character on the answer line below the register. To control the level of difficulty of the test problems, all groups were given identical test problems which were structured in accordance with their assigned experimental code. Selection of the test problems to be used, and the determination of whether they were to be encoded or decoded, was arrived at by random assignment.

Procedure

Each group was required to participate in two phases of the experiment. The two phases were administered on two consecutive days. Phase 1 was a one-hour group training session. Phase 2, administered on the following day, consisted of a group testing session, also approximately one hour long.

Task. Subjects were required to code character registers and address registers in the manner in which such coding occurs in actual field use. Programers in the field perform address register coding without the aid of a master code sheet since the task is compatible with established habits of handling decimal system numeric information. Character register coding however is frequently conducted by reference to master code sheets such as used in this study. The code sheet is necessary because many programers do not perform coding tasks with sufficient frequency to be able to recall character code configuration.

The format for the master code sheet for each code condition was identical. Letters of the alphabet were presented in alphabetic order in the first three rows, and the digits 0 to 9 were presented in numeric order in the fourth row. This format facilitated a systematic organization of the code sheet for both encoding and decoding tasks, regardless of the code schema. There was no apparent variation of this format which would improve the compatibility between the experimental task and any given master code sheet.

To perform encoding under any of the three conditions, the subject would locate the desired letter, digit, or symbol on the code sheet and make note of the circles containing an X. To decode under any of the three conditions, the subject systematically located the character by first determining which of the four rows to search in and then locating the character.

Training. Master code sheets were distributed to the members of each group during the training session. The code sheet corresponded to the unique code condition assigned to the group. The organization of the code sheet layout was explained to optimize search behavior habits. Test sheets identical to those to be used in the testing phase of the study were distributed to the subjects. An experimenter demonstrated to the subjects 10 representative decoding and 10 representative encoding problems. The demonstration was followed by practice trials by the subjects.

For each experimental condition, the same encoding and decoding problems were drawn from each of the four rows of the master code sheet and presented to the subjects as examples. One problem at a time was presented to each group and then checked for accuracy by both the subjects and two monitoring experimenters. Before continuing to the next problem, each subject was required to correct any errors. If the subject could not demonstrate to the experimenter that he could perform the correction unaided, the experimenter would again demonstrate the general coding procedure to him.

The criterion of success to which each subject was

trained was that of accurate unaided encoding and decoding of 24 consecutive coding problems. These problems were representative of all the types of problems to be presented in the test session. Because of the simplicity and highly redundant nature of the task, no subject under any condition required aid following the first five practice problems for any of the problem types (i.e., decoding a character register or address register, encoding a character register or address register). The need for aid was no more frequent under one condition than any other condition.

Upon completing the group training session the subjects were instructed that immediately prior to the testing session, a demonstration of coding would be repeated.

It was assumed that subjects would not rehearse the task in different amounts between the time of training and the time of testing for the following reasons: the task did not require memorization of stimulus code configuration, and therefore there was no need for rehearsal; because of the simplicity of the tasks the subjects reached complete mastery of the task early during the training session.

Testing. As indicated above a brief review of coding was presented to the subjects just prior to testing. Testing consisted of three trials with one test booklet per trial. Three sets of test booklets (1, 2, and 3) were used, each set consisting of different problems. To make a valid statistical test of the learning effect across trials, each trial required a random assignment of the test booklets to the subjects within a given code condition, with the restriction that each subject perform on a different booklet on each trial. Subjects were instructed to complete all problems in a test booklet as accurately and rapidly as possible, and to signal their completion by a raised hand. The time required by each subject was recorded by the experimenters.

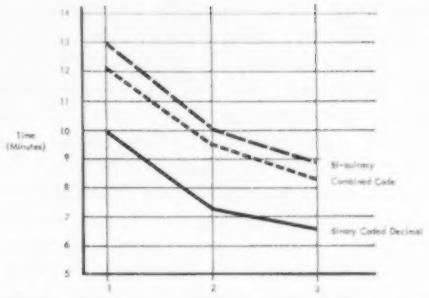


FIG. 3. Average time required to complete a test.

RESULTS

A 3 × 3 analysis of variance for repeated measurements on independent groups was performed for both the time and error criteria (Edwards, 1957).

Time Criterion Analysis

The 3 × 3 analysis of variance showed overall significant differences among code conditions and among trials ($p < .001$). A test of interaction between trials and code conditions was insignificant.

The results of the time criterion are presented graphically (see Figure 3). The results of the analysis of variance are summarized in Table 3.

To detect where the differences in mean time occurred between paired conditions, mul-

TABLE 3

SUMMARY OF ANALYSIS OF VARIANCE FOR TIME REQUIRED FOR TEST COMPLETION WITH THE 3 CODE CONDITIONS WITH 3 TRIALS FOR EACH GROUP

Source of variation	*SS*	*df*	*MS*	*F*
Between Codes I, II, III	214.5	2	107.3	19.5*
Between Subjects in Same Group	284.2	52	5.5	
Total between Subjects	498.7	54		
Between Trials A, B, C	395.2	2	197.6	286.3*
Codes × Trials	4.2	4	1.05	
Pooled				
Subjects × Trials	72.0	104	.69	
Totals within Subjects	471.4	110		
Total	970.1	164		

* $p < .01$.

tiple t tests were computed by the Scheffé (1953) method. It was found that the time required to perform the tasks using Code 2 was significantly less than the time required using either Code 1 or Code 3 ($p < .001$ and $p < .01$, respectively). The time required for Code 3 was less than that of Code 1, but the difference was not statistically significant.

Similar Scheffé tests were computed for the mean time required under each trial where the data were pooled for the three code conditions. The results showed that there was a significant reduction of the average time required with every successive trial ($p < .01$).

Analysis of Error Criterion

The 3×3 analysis of variance showed no over-all significant difference in the number of errors among the code conditions. However, there was a significant over-all difference in mean number of errors among trials ($p < .01$). The test of interaction between code conditions and trials was not statistically significant.

To detect where the differences in the mean number of errors occurred between paired trials, Scheffé sequential range tests were employed. It was found that Trials 2 and 3 had significantly fewer errors than the first trial ($p < .01$ and $p < .05$, respectively). The difference between Trials 2 and 3 was not statistically significant.

Additional analyses of the error criterion were performed. With the data of the three code conditions pooled, a t test for matched groups showed that there were significantly fewer errors made in the address register than in the character register ($p < .01$). It was also found that there were significantly fewer errors made in the encoding process than in the decoding process ($p < .01$).

No one condition had significantly more errors than another condition in either the character or address registers. In addition, no one condition had significantly more encoding or decoding errors than another condition.

The level of difficulty of the three sets of tests employed in the study were statistically checked. Neither the time nor error criterion revealed significant differences between the three tests.

Conclusions and Discussion

There were two major issues with which this study was concerned: (*a*) Which code system will permit the operator to utilize his code display registers with the most rapid rate of speed and result in the least number of errors? (*b*) With which code system can the operator improve his performance most rapidly?

With reference to *a*, the results indicate that the Binary Coded Decimal system (Code 2) permits coding performance to be carried out at a rate of speed which is more rapid than either of the other two codes. It requires 25% less time when compared to the Biquinary Code and 20% less time when compared to the Combined Code.

An inspection of time criterion data suggests that the greater proportion of differences in average time required between conditions was a function of character register coding. The rationale of this conclusion is that there was a significant difference between Code Condition 2 and Code Condition 3 although both code systems had the Binary Coded Decimal system used in the address register. It is hypothesized that the Biquinary Code systems require greater coding time as a function of the additional operator decision time necessary to determine whether a glowing numeric bit is to be read on the left or right side of that bit.

It is believed that the time advantage of the Binary Coded Decimal would be a permanent one. The rationale for this assumption is as follows: The coding task as performed at the data processing console by any one programer assigned to the system is not a high frequency task as compared to his major task of developing and writing new programs. Consequently, a great many programers do not practice sufficiently with the codes to memorize individual character code configurations. Coding performance in the field therefore is usually conducted with the aid of a master code sheet, as simulated in the experiment. Because of forgetting, it is hypothesized, that in the field, the level of coding performance and the performance profile from one console coding session to the next would

be very similar to the performance profiles of this experiment.

The analysis of the error criterion showed no significant statistical differences between the code conditions. This finding suggests that the rate of error would not be expected to differ for the three code conditions in field operations.

With reference to *b*, concerning the rate of performance improvement, the significant reduction in time required, and number of errors in successive trials, indicates that improved performance did take place. Under no code condition was learning rate more rapid than another condition. This conclusion is supported by the statistical finding that interaction between trials and code systems was insignificant. Therefore, the slopes of the plotted learning curves under each experimental code were near parallel for all practical purposes.

The results of this experiment have practical implications for selecting a code for a data processing system console. As stated earlier, the Binary Coded Decimal schema was the most efficient in terms of cost and electronic simplicity. The savings in operator coding time at the console, which would ensue from the use of this code, further justifies its use.

During program checking and debugging procedures at the console, the data processing system is nonproductive. Any reduction in system downtime constitutes an increase in system efficiency. There are many installations where new and complex programs must frequently be developed and checked for accuracy. It is in such installations that a 20–25% reduction in debugging time at the console can result in a practical increase in system productivity.

Summary

Three binary code schemas were evaluated experimentally in terms of operator coding performance. One of the three code schemas was to be selected for use on a data processing system console display. The criterion for the code evaluation was the operator efficiency measured in terms of speed and accuracy of coding console problems. An independent group of subjects was assigned to each of the three code conditions. One of the three code schemas facilitated a significant time saving of 20–25% as compared to the remaining two code schemas. There were no differences in rate of error between the three code conditions.

REFERENCES

Edwards, A. L. *Experimental design in psychological research*. New York: Rinehart, 1957.

Scheffé, H. A method for judging all contrasts in the analysis of variance. *Biometrika*, 1953, **40**, 87–104.

(Received November 16, 1960)

Journal of Applied Psychology
1961, Vol. 45, No. 6, 388–392

SUPERVISORY PROCEDURES AND WORK–TEAM PRODUCTIVITY[1]

J. S. KIDD AND R. T. CHRISTY

Ohio State University

In spite of the intensive concern of social psychologists with the general phenomenon of leadership, only fragmentary data are available concerning the interrrelations between the role specifications of the leader, the individual characteristics of the leader, the operational or task setting, and team productivity. Ever since the Hawthorne studies (Roethlisberger & Dickson, 1939) it has been recognized that the supervisor of a task oriented group or team can influence systematically its productive output. Lewin and his colleagues were able to demonstrate that the way in which leadership responsibilities were carried out was an important factor in determining many aspects of group behavior (Lewin, Lippit, & White, 1939). Again, in the context of corporate organizations, Shartle (1956) has pointed out that the personal characteristics of the executive are likely to predominate over the formal role requirements in determining the manner in which the leadership functions are fulfilled.

The present experiment was in no way intended to bring closure to the complex questions surrounding leadership role and team productivity. Its limited purposes were: to provide some insight into the relative influence of individuals vs. roles, to test the hypothesis that differential supervisory role specifications can have an effect upon team performance, to compare the effect of the different supervisory roles on more than one class of team performance criteria, and to develop some guidelines to facilitate the management of multiple team activities, especially those in which the team is to function in a man-machine system.

[1] This research was carried out in the Laboratory of Aviation Psychology and was supported in part by the United States Air Force under Contract No. AF 33(616)-6166, monitored by the Aerospace Medical Division. Permission is granted for reproduction, translation, publication, use, and disposal in whole and in part by or for the United States Government.

METHOD

Apparatus and Task Setting

The general task environment was provided by the simulation within the laboratory of a radar air traffic control center. The focus of observation was a three-man team consisting of two operators and a supervisor. This team was made responsible for the guidance of simulated aircraft through the preliminary phases of a landing approach. The task involved pickup and acknowledgment of aircraft entering a specified zone of responsibility, guidance of their flight course over a 50-mile approach route, altitude and airspeed adjustment prerequisite to actual landing, and positioning the aircraft for acceptance by a subsequent control agency for the final phase of the landing process. All task inputs were simulated.

The simulation was implemented by the specially developed OSU Electronic Air Traffic Control Simulator. This device, which is built around an analog computer, is capable of generating up to 30 dynamic aircraft targets and presenting them realistically to the radar controller via a cathode ray tube display. Direct manipulation of the "aircraft" is accomplished by college students trained to faithfully act the part of pilots. In addition to the visual display of aircraft position available to the controller, he is in direct auditory communication with the "pilots" under his jurisdiction through simulated radio channels.

Experimental Variables and Performance Measures

Three types of supervisor roles were compared in a context which included two levels of traffic input rate (load). The three supervisory roles were defined as follows:

Laissez-faire. In this condition the supervisor was instructed to act as a passive monitor of the on-going operation. He was available for consultation and was permitted to respond to direct questions asked by the controllers. He was responsible for keeping abreast of the situation, but did not intrude unless so requested by one of the controllers.

Active monitor. In this role, the supervisor acted as a "super-controller." He would observe carefully the progress of the operation in order to anticipate difficulties and detect controller errors. When either condition arose, the supervisor was to give directive instructions to the controller involved so as to permit the latter to correct the situation.

Direct participant. This role differed from the preceding primarily in that the supervisor was directed to take remedial actions by direct contact with the pilots. Thus, instead of requesting the controller to

carry out some corrective action, he could take the corrective action at his own discretion through a direct communication contact with the appropriate pilots. Coordination was automatic since the "channel" used by the supervisor was the same as that used by the controller, and any communication between supervisor and the pilots under a given controller's jurisdiction was overheard by that controller.

These three conditions can be regarded as defining a rough continuum of supervisory intervention. In all cases, however, the supervisor's actions are moderated by the presence and actions of his subordinates. Thus, even in the most extreme case, the supervisor does not have an opportunity to impose absolutistic control.

Two conditions of traffic input rate were utilized as a check on possible interaction effects. Rates of one aircraft entry every 45 seconds and one entry every 90 seconds were used.

Four performance measures were recorded. Mean percent delay was employed as the primary index of system efficiency. It reflects the rapidity with which aircraft are moved through the system and is defined as follows:

$$\frac{\sum_{1-n} \dfrac{\text{Actual} - \text{Hypothetical minimum flight time}}{\text{Hypothetical minimum flight time}}}{\text{Number of aircraft}}$$

A second measure was derived from the use, in the present study, of preprogramed pilot errors. Records were taken of the time lag between the occurrence of such a preprogramed error and its correction by the controller responsible. The programed errors were in the form of deviations from assigned heading. The third criterion consisted of a count of the number of controller errors made in positioning aircraft for the final portion of the landing process. The fourth measure was a count of the number of controller errors involving aircraft separation. If one aircraft "flew" within 2-miles lateral or 1,000-feet vertical distance of another aircraft, the controller was scored with a separation error.

TABLE 1

STATISTICAL DESIGN

Order of sessions	Conditions	Subjects (teams)					
		1	2	3	4	5	6
1	Supervisor	A	C	B	B	A	C
	Supervisory role	Laissez-faire	Laissez-faire	Direct participant	Direct participant	Active monitor	Active monitor
	Load	Low	High	High	Low	High	Low
2	Supervisor	A	C	B	B	A	C
	Supervisory role	Laissez-faire	Laissez-faire	Direct participant	Direct participant	Active monitor	Active monitor
	Load	High	Low	Low	High	Low	High
3	Supervisor	C	B	C	C	B	A
	Supervisory role	Direct participant	Active monitor	Active monitor	Laissez-faire	Laissez-faire	Direct participant
	Load	High	Low	High	High	Low	Low
4	Supervisor	C	B	C	C	B	A
	Supervisory role	Direct participant	Active monitor	Active monitor	Laissez-faire	Laissez-faire	Direct participant
	Load	Low	High	Low	Low	High	High
5	Supervisor	B	A	A	A	C	B
	Supervisory role	Active monitor	Direct participant	Laissez-faire	Active monitor	Direct paritcipant	Laissez-faire
	Load	High	Low	Low	Low	High	High
6	Supervisor	B	A	A	A	C	B
	Supervisory role	Active monitor	Direct participant	Laissez-faire	Active monitor	Direct participant	Laissez-faire
	Load	Low	High	High	High	Low	Low

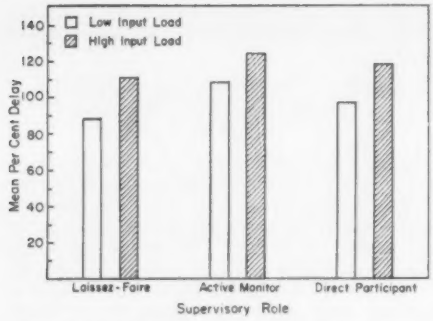


FIG. 1. Mean percent flight delay under various conditions of supervisory role and input load.

Subjects and Procedure

Specially trained college students participated both as controllers and as supervisors. Six two-man control teams were formed by random pairing from a 12-man sample. The three supervisors each had more than 2.5 years experience in the laboratory as controllers and had filled various leadership positions during this time. The test trials were arranged so that each supervisor had one opportunity to work with each of the teams. Thus, no control team experienced a supervisor in more than one role, although, in going from team to team, the supervisor changed roles. A modified greco-latin square design was used so as to balance load, role, and supervisor factors with practice effects. The design is detailed in Table 1.

A given experimental session represented a particular combination of one control team, a supervisor, and a role. Each session was 3.5 hours in duration and incorporated both input load conditions. Eighteen such sessions were required to complete the design requirements.

RESULTS

Three factors are of analytic concern: the influence of the supervisors as individuals on system performance, the influence of supervisory roles, and the influence of input load. All three factors and their interactions were compared in terms of the mean percent delay criteria. The remaining criteria were employed only with respect to the comparison of the effects of supervisor roles.

The first step in analysis was an over-all comparison by analysis of variance of the three factors using the delay criteria. This was done primarily to assess the comparative variance contributed by supervisors as individuals vs. the variance contributed by the role assignments and as a check on interactions rather than as a direct comparison of the conditions within a given factor. The influence of supervisors as individuals yielded a mean square variance 3.7 times greater than that derived from the role factor. The interactions of role and supervisor and of role and load were statistically negligible. The influence of these factors may be seen directly by reference to Figures 1 and 2. The relationship between role and supervisor is of potential interest since the best of the three supervisors is also the least variable insofar as this can be assessed by the performance of his crews.

Since the effect of interaction between the experimental variables was relatively minor, direct comparison of conditions within factors was undertaken by nonparametric methods. Table 2 presents the comparison of the three supervisory roles in terms of all four criteria of system performance. In all cases the analysis is carried out between pairs of conditions, since they represent discrete categories rather than a continuous dimension. It will be noted that the laissez-faire role is superior according to the delay criterion but that active monitoring is superior with respect to detection of programed pilot errors and with respect to avoidance of positioning errors (in the final stages of a flight), while the direct participant role is superior in the avoidance of separation errors.

The input load factor evaluated alone was highly significant. Under low load conditions,

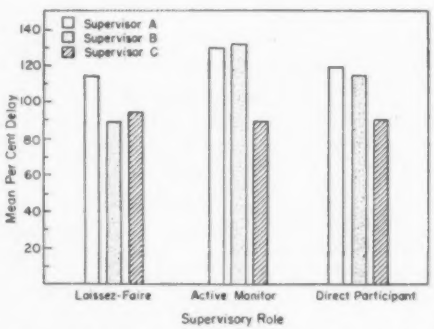


FIG. 2. Mean percent flight delay with different individual supervisors taking various supervisory roles.

TABLE 2

THE INFLUENCE OF SUPERVISORY ROLE ON VARIOUS MEASURES OF SYSTEM PERFORMANCE

Criteria	Role designation			Paired tests[a]		
	Laissez-faire	Active monitor	Direct participant	Laissez-faire vs. active monitor	Laissez-faire vs. direct participant	Active monitor vs. direct participant
Mean percent delay	99.2	116.5	107.4	$d_1 + d_2 > 0$ $p = .062$	*ns*	*ns*
Pilot error detection lag	92.7	82.2	97.8	*ns*	*ns*	$d_1 + d_2 > 0$ $p = .062$
GCA go-arounds per 100 landings	4.7	2.6	2.9	$d_1 + d_2 > 0$ $p = .062$	$d_1 > 0$ $p = .031$	*ns*
Separation errors per 100 landings	4.5	3.9	2.9	*ns*	$d_1 + d_2 > 0$ $p = .062$	*ns*

[a] The Walsh test was used throughout (Siegel, 1956).

the average percent delay was 97.8, while under high load conditions it was 117.6. By the Walsh test, such a difference has a probability of chance occurrence of .031.

DISCUSSION

The relatively small sample size employed in the present experiment limits, to some degree, the extent to which generalizations can be made. The major virtue of the experiment is perhaps in the fact that problems usually susceptible only to field investigation were brought into the controlled laboratory environment.

Thus, some quantitative validity can be given to the proposition that individual supervisor characteristics are more influential on work-team performance than are the various standard styles of carrying out the supervisory function. The problem remains, however, as to just exactly what are such personal characteristics. Although there are supervisor selection techniques available, further vigorous study of the problem should be given added incentive by the present findings.

The comparison of the supervisory roles, per se, leads to the proposition that there is a distinct trade-off effect between performance criteria. Thus, the laissez-faire role with its nonintervention feature allows the controllers to concentrate on maintaining a rapid flow through the system. The controller apparently accomplishes this particular end with a certain recklessness since his error scores are relatively high. When the supervisor expresses his role more aggressively, errors are reduced, but at a cost in the rapidity of flow. The management implication seems clear. Recommended supervisory procedure depends on the importance of speed vs. accuracy in the task at hand. Apparently, high levels of accuracy demand more overt intervention by the supervisor. When both are equally important, a moderate level of intervention would seem most appropriate.

This conclusion should also be tempered by the recognition that novice controllers will require more active participation on the part of their supervisors so that minimal performance levels are consistently assured. Thus, adoption of the laissez-faire approach would depend on both the nature of the task and the level of achieved skill on the part of the operators.

In the present study, the supervised controllers were informally interviewed subsequent to their participation in the experiment. Their comments were fairly consistent: they preferred the laissez-faire condition. Their response to the other conditions changed with time. At first, they were somewhat resentful to the supervisor's "interference," but as time passed they were more inclined to perceive the supervisor as at least well intentioned. Therefore, an effective supervisor might be one who promptly made his intentions manifest to his team.

Summary

A complex task setting provided by the simulation of a radar air traffic control system was the context employed to evaluate the extent to which work-team productivity is modifiable as a consequence of different supervisors and supervisory procedures under different task loads. Six two-man teams were observed. Each team worked under each of three supervisors. The supervisors shifted from team to team. Three techniques were adapted alternately: laissez-faire, active monitoring, and direct participation.

It was observed that the individual supervisor was a more consistent influence on performance than the particular role he employed. The effect of role per se was significant but interpretable only in light of the particular performance criterion used. For example, processing speed was greatest under laissez-faire conditions, while error avoidance was superior under the conditions which required more overt supervisory activity. Task load was not found to be an interactive factor.

REFERENCES

Lewin, K., Lippit, R., & White, R. K. Patterns of aggressive behavior in experimentally created "social climates." *J. soc. Psychol.*, 1939, **10**, 271–299.

Roethlisberger, F. J., & Dickson, W. J. *Management and the worker.* Cambridge: Harvard Univer. Press, 1939.

Shartle, C. L. *Executive performance and leadership.* Englewood Cliffs, N. J.: Prentice-Hall, 1956.

Siegel, S. *Nonparametric statistics for the behavioral sciences.* New York: McGraw-Hill, 1956.

(Received November 25, 1960)

Journal of Applied Psychology
1961, Vol. 45, No. 6, 393–401

BALES' INTERACTION PROCESS ANALYSIS OF PERSONNEL SELECTION INTERVIEWS[1]

DANIEL SYDIAHA

University of Saskatchewan

The purpose of this investigation was to examine whether Bales' (1950) categories of interaction process analysis were related to decisions made by personnel selection interviewers.

It has been generally recognized that decisons made by interviewers are notoriously unreliable, in the sense that assessments made by two or more interviewers of the same applicants tend to differ markedly. The question as to the sources of this unreliability, however, has not been dealt with in the literature, except for casual or speculative mention of various situational factors such as unsystematic questioning or interviewer bias. In the absence of empirically supported generalizations about sources of error, the practising interviewer is left with little more than a feeling of disdain toward a technique which is generally believed to be inadequate.

The hypothesis advanced in the work reported here is that decisions made by interviewers in personnel selection are functionally related to measurable dimensions of interview process and content. The point of view adopted is essentially Brunswick's notion of probabilistic functionalism (Brunswick, 1956). While this is admittedly a simple hypothesis, it is nonetheless not a naive one, since the more common assumption about interviewing is that interview process and content are unsystematic and chaotic. The approach taken has been frankly empirical, implying no strong predisposition as to the nature of interview process, but rather an attempt is made at developing three sets of variables which have some plausibility in being functionally related to decisions. These variables are: (*a*) information bias, reported in a previous paper (Sydiaha, 1959); (*b*) Bales' interaction process analysis, reported here; and (*c*) interviewer empathy, to be reported in a subsequent paper.

In casting about for measures which might be expected to show some uniformity in interview situations, Bales' system of interaction process analysis was selected as a promising method because of Bales' success in demonstrating the existence of certain uniformities in group process (see, for example, Bales, 1950; and articles in Hare, Borgatta, & Bales, 1955; Maccoby, Newcomb, & Hartley, 1958; Swanson, Newcomb, & Hartley, 1952). These uniformities in interaction patterns have lead to some useful theoretical statements on the nature of group process. Other more specific differentiations have been made of more direct relevance to decision making in interviewing, such as the differences in interaction profiles between (*a*) successful and unsuccessful groups, as described by group participants; and (*b*) group leaders who are task oriented (the "idea-man") and person oriented (the "social-emotional specialist").

The interaction categories are described elsewhere (Bales, 1950).[2] Briefly it may be noted that a distinction is made between acts which are classified in "task" categories (4, 5, 6, 7, 8, and 9) and "social-emotional" categories (1, 2, 3, and 10, 11, 12). Within these two broad divisions, there is a further division of task acts into questions (Categories 7, 8, 9) and attempted answers (4, 5, 6) while so-

[1] Financed by Defence Research Board of Canada, Grant No. 9435-53 to E. C. Webster, McGill University. Webster directed the research and Allen Clark and John Kenyon assisted with computations. The author gratefully acknowledges the assistance of all interviewers who assisted in providing information, and in particular, members of the Canadian Army Personnel Selection Service.

[2] The 12 Bales categories are briefly designated as follows: 1. shows solidarity, 2. shows tension release, 3. agrees, 4. gives suggestions, 5. gives opinion, 6. gives information, 7. asks for information, 8. asks for opinion, 9. asks for suggestion, 10. disagrees, 11. shows tension, 12. shows antagonism. Two additional categories used in this study were: 7A. asks for information so as to require "yes" or "no" responses, 6A. "yes" or "no" responses to 7A questions.

cial-emotional acts divide into positive (1, 2, 3) and negative (10, 11, 12) acts. In addition to these 12 categories proposed by Bales, two other categories were added for this study, namely, Categories 6A and 7A, these being subcategories of 6 and 7. Category 7A acts were those usually scored in Category 7 ("asks for information, orientation, etc.") but which are framed in such a way as to require a simple "yes" or "no" response. Similarly Category 6A acts were those usually scored in Category 6 ("gives information, orientation, etc.") but they are expressed as simply "yes" or "no." The purpose of adding these categories was to test the consequences of what might be termed "directive" questioning in which the applicant has little or no control over the information communicated but can only reply "yes" or "no" to questions put to him. This pattern of questioning is in contrast to a more nondirective situation in which questions are open ended, the applicant is given considerable freedom of choice in his responses and, in a sense, is able to exert some control over the interview. It seemed apparent, from informal observations made during preliminary stages of the study, that directive questioning tended to have one of two consequences: either an atmosphere was created of the interviewer cross-examining the applicant, or it signified that the interviewer had lost control of the situation in that he had exhausted his supply of questions and was frantically searching for things to talk about. In either case, it seemed apparent that the situation was an uncomfortable one for both applicant and interviewer, and tended to be associated with rejection of the application.

The following hypotheses were proposed, consistent with the above discussion:

H_1: Comparing interviews in which applicants are accepted with those in which applicants are rejected: (*a*) interaction scores are higher for Categories 1, 2, 3, 4, 5, 6, and (*b*) lower for Categories 7, 8, 9, 10, 11, 12, 6A, and 7A.

H_2: (Consistent with the notion of inter-interviewer error) Differences in interaction scores between accept and reject cases are greater for some interviewers than for others.

Two studies were executed, designated Studies I and II.

Study I

Procedure

Training in the use of Bales' interaction process analysis. The method of scoring described by Bales (1950) was adhered to as closely as possible. Sound recordings of interviews were obtained by the experimenter (the writer) and scoring of these interviews was repeated until two consecutive scorings of the same interviews gave comparable results. Comparability was tested by means of the chi square test as described in Bales (1950). To check on comparability of results, another scorer (A. Clark) received some training in the scoring method, and marked discrepancies between the two scorers were discussed, and criteria for agreement decided upon, in light of Bales' instructions. As an added check on scoring, the experimenter visited Bales at Harvard and discussed scoring problems with him.[3]

Training was carried out with a sample of 25 interviews. Upon completion of training, a sample of 67 interviews was scored twice to test the reproducibility of scoring. Time intervals between consecutive scorings varied from 6 hours to 3 weeks. For each scoring category, a correlation (Pearson product-moment) was calculated between the two sets of scores. The results are shown in Table 1. The reproducibility of scores shown in Table 1 is high although the correlations are undoubtedly elevated by the recollection of the first scoring on the second occasion.

Subjects. One hundred fourteen interviews were obtained from nine interviewers, with a range of 9–24 per interviewer. All applicants and interviewers were male. The interviewers included six Canadian Army personnel officers, one Royal Canadian Air Force career counselor, one industrial employment interviewer, and a psychologist engaged in private consulting practice. Five of the Army officers examined recruit applicants and the sixth examined

[3] Bales' assistance and interest in this study are gratefully acknowledged.

TABLE 1

Correlations (Pearson Product-Moment) between Two Repeated Scorings of Recorded Interviews ($N = 67$)

Category	*r*	Category	*r*	Category	*r*
1	.89	6	.97	9	—[a]
2	.70	6A	.95	10	.71
3	.98	7	.98	11	.96
4	.84	7A	.98	12	.90
5	.96	8	.87		

[a] Acts in Category 9 occurred very infrequently and were not analyzed.

officer applicants. The Air Force interviewer examined applicants for both air crew and administrative positions. The industrial interviewer examined principally junior office and clerical applicants. The subjects obtained from the psychologist varied widely in terms of qualifications, from industrial apprentices to senior executives, and for purposes of this study, the sample was divided into "Junior" and "Senior" groups. The number of interviews in the 10 groups varied from 9 to 15.

The selection of the samples of interviewers and applicants is in no sense to be considered representative of any population of selection interviews regularly conducted, but rather, merely a sample of accessible interviews obtained from interviewers who could be persuaded to contribute data for the study.

Conditions of interviewing. The interviews were "natural" except that (*a*) in most cases (approximately 80%) the experimenter was present in the interview room as an observer, and (*b*) except for five interviews, all interviews were sound recorded. Concealment of the sound recorder was left to the discretion of the interviewer, and in about 60% of the cases the applicant did not know that there was a recorder in the room.

Upon completion of each interview, the interviewer recorded his recommendation as to whether the applicant should be accepted, rejected, or whether the case was marginal or doubtful. Because the interviews were selected more or less at random from the stream of applicants examined, it was not possible to ensure that the numbers of cases would be the same for all three categories or for all interviewers. In some instances only a single case was obtained for a particular decision category. Since no estimate of variance could be made in this instance, it was decided to combine the "marginal" and "reject" categories, thereby including all cases obtained for the analysis. Admittedly, this combining of judgment categories produced a questionable division of cases since many of the marginal cases were in fact accepted for employment although for purposes of analysis they were classified as having been rejected. In defence of this procedure, it is to be noted that the criteria for employment were necessarily not uniform because of differing employment settings in which the interviewers worked. Furthermore, Study I was a preliminary one only, and was designed to test whether differences could be found between *any* decision categories. Study II made use of consistent as well as working criteria of acceptance-rejection.

Analysis of data. Interaction scores for each interview were transformed into percentage figures in order to equate for interview duration, and these percentage figures for each category were analyzed by complex analysis of variance with double classification. In other words, a separate analysis of variance table was prepared for each of the 13 categories. (Category 9 was excluded from analysis because there were too few acts in that category to warrant analysis.) In each analysis of variance table, rows represented decision categories, columns represented separate interviewer-groups, and the interactions represented interviewer-decision category interactions. Since all frequencies were unequal, it was necessary to make adjustments by the method of expected equal frequencies, as described in Ferguson (1959).

TABLE 2

MEAN SQUARES OF ANALYSIS OF VARIANCE COMPONENTS FOR SCORING CATEGORIES

(Study I, $N = 114$)

Category	R ($df = 1$)	C ($df = 9$)	RC ($df = 9$)	*Ss* ($df = 94$)
1	11.50**	25.74**	1.03	1.60
2	13.90	97.95**	5.12	3.53
3	45.98**	251.09**	8.30	8.98
4	2.60	2.70	4.50**	.94
5	11.24	30.82	17.31	21.19
6	69.80	242.48	17.53	41.45
6A	3.07	29.99*	4.96	4.01
7	8.98	252.10**	16.99	12.82
7A	10.68	68.83**	10.86	48.24
8	15.40	55.83*	13.44*	6.64
10	1.22*	.17	.16	.18
11	.01	17.54**	3.24	2.17
12	13.76	7.60	3.37**	1.20
Duration of interview (minutes)	.23	1979.26**	82.70	91.67
1 + 2 + 3	190.33*	469.26**	18.34	19.42
4 + 5 + 6	.01	139.71*	15.28**	2.42
7 + 8 + 9	33.29	488.10**	33.10	17.78
10 + 11 + 12	49.34	26.66	11.70**	3.46

Note.—C = interviewer samples. R = decision categories. *Ss* = individual interviews.
* *F* ratio statistically significant at 5% level.
** *F* ratio statistically significant at 1% level.

Other variables analyzed were duration of interview, and each of the following combined scores: $1 + 2 + 3$, $4 + 5 + 6$, $7 + 8 + 9$, $10 + 11 + 12$. Mean squares for row, column, interaction, and individual interview components are shown in Table 2. *F* ratios were calculated for all main effects and interactions. Following McNemar (1955), the denominators used for the *F* ratio were: individual interviewers mean square to test for the significance of interactions, and interactions mean square to test for the significance of main effects.

Results

Decision category differences. Significant differences between decision categories were obtained for Categories 1 (shows support), 3 (agrees), $1 + 2 + 3$ (positive social emotional acts), and 10 (disagrees). Mean category scores, not presented in this paper, were consistent with expectation: "accept" interviews were characterized by high scores in positive social emotional categories, and by low scores in Category 10.

It should be noted that of these four categories, three also showed significant differences among interviewers (Categories 1, 3, and $1 + 2 + 3$). For these categories, then,

mean scores are not comparable for all interviewers.

Interviewer-decision category interactions. Significant interactions were obtained for Categories 4 (gives direction), 8 (asks for opinion), 12 (shows antagonism), 4 + 5 + 6 (problem solving attempts), and 10 + 11 + 12 (negative social emotional acts). Mean category scores were again consistent with expectation, although there were some reversals. In general, "accept" interviews tended to be characterized by high scores in Categories 4 and 4 + 5 + 6, and by low scores in Categories 8, 12, and 10 + 11 + 12.

These results imply that decision making is related to frequency of occurrence of acts in these categories, but not for all interviewers. These categories suggest sources of interview process which account for interinterview error.

Of the five categories, two also showed significant differences among interviewers (Categories 8 and 4 + 5 + 6). This implies noncomparability of scores among interviewers, although these findings have no special significance since, in any case, all significant interactions imply noncomparability of scores.

Interviewer differences. Significant differences between interviewers were obtained for Categories 2 (shows satisfaction), 6A (gives information in "yes" or "no" terms), 7 (asks for information), 7A (asks for information with questions to be answered as "yes" or "no"), 11 (shows tension), duration of interview, and 7 + 8 + 9 (questions). These results are of no special significance, however, since they are not related to the decision categories.

Discussion

Hypotheses 1 and 2 were confirmed: some of the Bales scoring categories were correlated with decisions made by personnel selection interviewers. In general favorable decisions were characterized by high scores in problem solving attempts and positive social emotional acts, and by low scores in questions and negative social-emotional acts. However, the differences between decisions were not uniform among interviewers.

In general, the results of this preliminary study showed promise of depicting something about the nature of interview process, with regard to uniformities both *specific* and *common* to interviewers.

Demonstrating the presence of a relationship between categories and decisions leaves unanswered the question of the magnitude of the relationship. Because the interview groups in Study I were too small to permit the calculation of correlation coefficients between variables, Study II was conducted with larger numbers of interviews per interviewer to permit such estimates of degree of relationship. Furthermore, all cases were obtained from one organization (the Canadian Army) in order to standardize the criteria of acceptance-rejection used.

Study II was also set up to examine the importance of interviewer and applicant scores considered separately. Up to this point, each category score had included the combined acts of both applicant and interviewer and the results considered in terms of the relevance of interaction scores to decision making, regardless of which participant contributed acts to the conversation. Separate participant scores, on the other hand, are related to questions of individual role performance as it relates to decision making in the interview.

Some evidence in support of distinctive applicant and interviewer profiles was obtained from the data in Study I, but it was not included in Table 2 because many of the category scores were found to be heterogeneous, thus rendering use of the F ratio somewhat tenuous. Analysis of the data using Friedman's nonparametric analysis of variance (see Ferguson, 1959) yielded differences between decision categories for some scores, but since the statistical procedure does not assess interaction effects, the data were not considered comparable to those shown in Table 2, and are omitted from this paper for that reason.

Study II

Procedure

Details of the study are described elsewhere (Sydiaha, 1959). Each of eight Canadian Army Regular Force personnel officers interviewed from 13 to 50 Regular Force applicants for the Canadian Army. Total N for the project was 256, but data for Bales' data were missing in 6 cases. Interviews were conducted under usual Army circumstances except for additional research material requested for the stud-

ies herein reported, and the interviews were sound recorded. The recording machines were in plain view of all applicants, and they were informed that the recording was to be used for research purposes. Only the applicant and the interviewer were present in the room during each interview. Each case was assessed by only one officer, who classified the applicant as either "accepted" or "rejected." Rejected cases included 20 applicants recommended for assessment at a later date and 7 applicants referred for psychiatric assessment. Accepted cases included 29 applicants considered marginally suitable for Army service.

The cases were divided into three separate groups to permit cross-validation of findings. Cases provided by Officers A, B, F, and G ($N = 37$, 50, 50, 41) were randomly assigned to a criterion group ($N = 88$) and a holdout group ($N = 90$). The remaining cases, i.e., those provided by Officers C, D, E, and H ($N = 13$, 23, 18, 18) made up a second holdout group ($N = 72$).

Scoring was performed from disc sound recordings by the experimenter (the writer) as in Study I. For each interview, interaction scores were converted to percentages for the applicant and interviewer scores separately in order to partial out the effects of the total amount spoken by each of the participants. (These were treated as separate scores.) The general statistical procedure followed was to combine all scores mathematically in such a way as to maximize the correlation between combined scores and acceptance-rejection for the criterion group and the scoring key was then applied to the holdout group to test the relation between scores and acceptance-rejection. Three such combined scores (referred to henceforth as "indices") were developed: the "Applicant Index" (made up of applicant scores), the "Interviewer Index" (made up of interviewer scores), and the "Interview Duration Index" (made up of the total number of acts scored for each of the two participants and the duration of the interview in minutes).

The discriminant function was used to make up the Interview Duration Index. This procedure was not felt to be justified for making up the other two indices, however, because of the large number of variables involved. Guilford (1954), for example, has pointed out the inadvisability of multiple regression analysis for combining test scores when their number is high. Consequently these indices were made up by the standard error method of weighting, described by Gulliksen (1950) whereby scores are weighted as a function of the standard deviation and the reliability coefficient. Variables which correlated negatively with acceptance-rejection (for the criterion group) were assigned negative weights. After all three indices had been calculated, they were combined into a "Total Index" by means of a discriminant function.

As was indicated in the results of Study I above, interaction scores are not uniform for all interviewers. To assess the consequences of such interviewer differences in Study II, two separate criteria of item selection were used, which, in turn, lead to the development of two sets of scoring keys for combining scores, and two sets of indices. The two criteria were: acceptance-rejection for individual Officer Samples A, B, F, G; acceptance-rejection for the combined Officer Sample A + B + F + G. The reasoning involved in the use of these two criteria was that, on the one hand, items correlated with acceptance-rejection for individual officers would reflect criterion variance *specific* to each officer, whereas items correlated with acceptance-rejection for the combined officer sample would reflect criterion variance *common* to all officers. Depending upon the relative proportion of specific- and common-criterion variance present in the data, the following results might be expected to occur: (*a*) If criterion variance common to all officers were zero, then correlations between indices and acceptance rejection would be zero for the combined holdout samples (A + B + F + G) and (C + D + E + H). In other words, the magnitude of such combined sample correlations would be indicative of criterion variance common to all officers. (*b*) If criterion variance specific to officers were zero, then correlations between indices and acceptance-rejection would be the same, regardless of whether indices were based upon individually based or group based scoring keys. In other words, correlation differences between indices derived from these two scoring keys would be indicative of criterion variance associated with specific officers.

Consistent with such reasoning, the procedures adopted were as follows:

Individual scoring keys: All variables were correlated with acceptance-rejection for each of the four officer criterion groups, namely, A, B, F, G. The scoring weights derived were applied to each of the corresponding officer holdout groups. This procedure was intended to *maximize* criterion variance specific to the four officers.

Group scoring keys: All variables were converted to standard score form, based upon the distribution of scores obtained for each officer. All variables were then correlated with acceptance-rejection for the combined criterion group, namely, A + B + F + G. The scoring weights derived were applied to the two holdout groups, namely, A + B + F + G and C + D + E + H. This procedure was intended to *minimize* criterion variance specific to the four officers.

Results

Group scoring keys.

Applicant indices: Correlations between applicant indices and acceptance-rejection were .31 and .23 for the two holdout groups (see Column 4, Table 3). These correlations are statistically different from zero ($p < .005$). (Point-biserial correlations were tested for statistical significance by the method described in Ferguson, 1959.) The difference between these correlations was not significant, thus making it possible to calculate a correla-

TABLE 3

CORRELATION (POINT-BISERIAL) BETWEEN ACCEPTANCE-REJECTION AND THREE BALES' INDICES, USING INDIVIDUAL SCORING KEYS AND GROUP SCORING KEYS
(Study II)

1	2	3	4	5	6	7	8	9	10
		Applicant Index		Interviewer Index		Interviewer Duration Index		Three indices combined[a]	
Sample	*N*	ISK	GSK	ISK	GSK	ISK	GSK	ISK	GSK
A	19	.22	.31	.39	.42	−.05	−.07	.45	.43
B	25	.40	.32	.70	.55	.28	.22	.68	.50
C	13		—[b]						
D	23		.20		.43		.32		.42
E	18		.24		.24		.18		.39
F	25	.35	.12	.76	.52	−.33	.14	.38	.41
G	21	.27	.46	.47	.15	−.05	.00	.24	.28
H	18		.49		.52		.35		.68
A + B + F + G	90	.32[c]	.31	.62[c]	.34	−.04[c]	.08	.46[c]	.37
C + D + E + H	72		.23		.25		.19		.32
All cases	162		.28		.29		.12		.35

[a] Indices were combined via discriminant function.
[b] Officer C had no rejected applicants in his sample.
[c] Mean correlations.

tion based on all cases, which was .28. Differences among the seven officer-sample correlations shown in Column 4 of Table 3 were not significant ($p > .9$). (Correlation differences among officer samples in this paper were tested by the method described in Snedecor, 1956.)

Interviewer indices: Correlations between interviewer indices and acceptance-rejection were .34 and .25 for the two holdout groups (see Column 6, Table 3). These correlations are statistically different from zero ($p < .005$ and .025, respectively). The difference between these correlations was nonsignificant, and the correlation coefficient for all cases combined was .29. Differences among the seven officer-sample correlations shown in Column 6 of Table 3 were not significant ($p > .8$).

Interview duration indices: Correlations between the interview duration indices and acceptance-rejection were .08 and .19 for the two holdout groups (see Column 8, Table 3). These correlations are not statistically different from zero, as determined by the test of the significance of multiple correlation, described by Ferguson (1959).

Three indices combined: Correlations between the three indices combined and acceptance-rejection were .37 and .32 for the two holdout groups (see Column 10, Table 3). Using the test for the statistical significance of multiple correlation, only the first correlation is different from zero ($p < .01$). However the two correlations are not significantly different, and the multiple correlation for all cases combined is .35, which is nonzero ($p < .01$). Differences among the seven officer-sample correlations shown in Column 10 of Table 3 were not significant ($p > .9$).

The results tend to support Expectation *a:* With the exception of the interview duration indices, all measures were correlated with acceptance-rejection. The scoring procedures used were successful in eliminating interinterviewer differences in interaction scores, while at the same time, demonstrating consistent interaction process common to all interviewers.

Individual scoring keys.

Applicant indices: The average correlation between applicant indices and acceptance-rejection was .32 compared with .31 for indices based on the group scoring keys (see Columns 3 and 8, Table 3).

Interviewer indices: The average correlation between interviewer indices and acceptance-rejection was .62 compared with .34 for indices based on group scoring keys (see Columns 5 and 6, Table 3). The question as to whether the difference between these correlations is significant could not be answered directly and a simplifying assumption was made in order to obtain an estimate of statistical significance. The reason that a direct test of significance could not be made in this instance was that a correlation between the two sets of indices was required in order to test the correlation differences since both correlations were based on the same subjects (see Ferguson, 1959). Such a correlation between indices was not feasible in this case since there were wide differences in mean scores among the four officer samples based on the individual scoring keys.

The simplifying assumption made was that the correlation between the two indices was zero, which represents a conservative estimate of the true state of affairs. Using this assumption, the difference between the two correlations (.62 vs. .34) was statistically significant ($p < .05$).

Considering individual officer samples in Columns 5 and 6 in Table 3, correlation differences were not significant in any case, although the small sample sizes involved would militate against demonstrating any significant results.

Interview duration indices: The average correlation between interview duration indices and acceptance-rejection was $-.04$ compared with .08 for indices based on group scoring keys (see Columns 7 and 8, Table 3).

Three indices combined: The average correlation between the three indices combined and acceptance-rejection was .46 compared with .37 for the indices based on the group scoring keys (see Columns 9 and 10, Table 3).

The results obtained from the interviewer scores tend to support H_2: Indices based on individual keys account for more criterion variance than do indices based on group keys. Applicant and interview duration indices do not yield different results when comparing individual with group scoring keys.

Discussion

Hypothesis 1 was confirmed: Bales' interaction process analysis scores were correlated with decisions made by personnel selection interviewers. The correlations obtained were modest in magnitude, although they were uniform across all interviewers, provided the data were transformed into standard scores. Taking the correlation coefficient of .35 based upon all cases and combining all categories it may be concluded that approximately 12% of the criterion variance is accounted for by interaction score variance common to all interviewers.

Confirmation of Hypothesis 2 was confined to interviewer conversation only: interviewer differences in interaction scores between accept and reject cases were found only for interviewer indices, not for the applicant and interview duration indices. Taking .62 as being the correlation between interviewer indices and acceptance-rejection we may estimate the criterion variance thus accounted for to be 38%, which presumably includes both common and specific variance. By subtracting the estimated common variance (12%), mentioned above, it would appear that criterion variance specific to interviewers is of the order of 26%.

It should be understood that these variance estimates are, at best, very rough, since the correlations shown in Table 3 are based on very small samples, and there may undoubtedly be differences among interviewers in the magnitude of criterion variance accounted for by specific interaction score variance. However, as a first approximation, it may be concluded from this study that such specific interaction score variance accounts for inter-interviewer error to a significant degree.

It should be pointed out that the occurrence of such interviewer differences precludes the making of definitive statements as to "typical" interaction scores in employment interviews. While mean scores in Study II tended to be consistent with the results of Study I ("accept" interviews tended to be characterized by high scores in positive social-emotional, and problem solving categories, and by low scores in questions and negative social-emotional categories) these

trends were somewhat incidental. Mean scores varied considerably among decision categories as well as among interviewers, and reversals in the direction of mean differences were common.

The question remains as to the implications of these findings for interview techniques. Clearly, this study has dealt with the assessment of the concomitants of decision making rather than with manipulable antecedents, such that any interpretation must be considered speculative, pending confirmation by experimental manipulation of relevant circumstances. At least two interpretations of the results are apparent:

1. Interaction process scores may reflect valid dimensions of job performance such as interpersonal skills. Interinterviewer differences in interaction process scores would correspond to differences in sensitivity to such skills. (A follow-up is now under way in which interaction process scores will be correlated with criterion measures based on the inductees' performance records following 3 years regular Army service.)

2. Depending upon whether interaction process scores are correlated with documentary and other descriptive information presented during the interview, these results would throw some light on the possible importance of nonrational processes in decision making. To illustrate this point, let it be assumed that interaction process scores and descriptive information are uncorrelated. Then the fact that interaction process scores are correlated with acceptance-rejection would imply that the interviewer is sensitive to influences which are beyond the realm of data gathering, and may possibly overlook information in reaching his decision, by reacting to a "favourable atmosphere." Obviously, if the existence of such an atmosphere is irrelevant to his task as formerly defined, then the interviewer may be subject to error in his decisions. Particularly if it could be shown that applicant subjects could establish a favorable atmosphere by "role playing" certain interaction categories, these results would serve to demonstrate some of the pitfalls of interviewing technique.

On the other hand, if interaction process scores were found to be correlated with descriptive information, then the results obtained in this study would be of little consequence, since this would imply that interaction scores merely reflected the consequences of favorable or unfavorable information (whichever the case might be) divulged during the interview.

The data presented here cannot settle the issue as to which interpretation is more plausible. One appropriate test of the alternatives would involve having experimental applicant subjects attempt to "role play" or "fake" certain interaction categories in order to bring about favorable decisions on the part of interviewers. It may be worth emphasizing that the data discussed here are entirely consistent with the notion that an applicant *can* fake his way to a favorable decision. (The data are also consistent with the notion that interviewers respond to such faking in terms of different interaction profiles.) But such interpretations must wait on further investigation.

Summary

Samples of interview conversation (personnel selection interviews) were analyzed according to Bales' interaction process analysis. Scores obtained were correlated with decisions made by interviewers about whether applicants were recommended for acceptance or rejection.

In Study I, data were obtained from nine interviewers (10 groups of cases ranging from 9 to 15) and scores were submitted to analysis of variance procedure. Some significant differences in interaction process scores were found between decision categories, and some interviewer-decision interactions were also statistically significant.

In Study II, data were obtained from eight Canadian Army personnel officers who interviewed from 13 to 50 applicants for the Army. Correlations between interaction categories and acceptance-rejection were calculated using item analysis and cross-validation procedures.

The results indicated that: (*a*) Bales interaction process categories are correlated with acceptance-rejection, the correlation being low but consistent for transformed data; (*b*) interinterviewer differences in interaction process are confined to interviewer conversation only.

The implications of the results are discussed, in terms of the possible significance of applicant faking of favorable interactions.

REFERENCES

BALES, R. F. *Interaction process analysis: A method for the study of small groups.* Cambridge: Addison-Wesley, 1950.

BRUNSWICK, E. *Perception and representative design of psychological experiments.* (2nd ed.) Berkeley: Univer. California Press, 1956.

FERGUSON, G. A. *Statistical analysis in psychology and education.* New York: McGraw-Hill, 1959.

GUILFORD, J. P. *Psychometric methods.* (2nd ed.) New York: McGraw-Hill, 1954.

GULLIKSEN, H. *Theory of mental tests.* New York: Wiley, 1950.

HARE, A. P., BORGATTA, E. E., & BALES, R. F. (Eds.) *Small groups: Studies in social interaction.* New York: Knopf, 1955.

MACCOBY, ELEANOR E., NEWCOMB, T. M., & HARTLEY, E. L. (Eds.) *Readings in social psychology.* (3rd ed.) New York: Holt, 1958.

MCNEMAR, Q. *Psychological statistics.* (2nd ed.) New York: Wiley, 1955.

SNEDECOR, G. W. *Statistical methods.* (5th ed.) Ames, Ia.: Iowa State Coll. Press, 1956.

SWANSON, G. E., NEWCOMB, T. M., & HARTLEY, E. L. (Eds.) *Readings in social psychology.* (2nd ed.) New York: Holt, 1952.

SYDIAHA, D. On the equivalence of clinical and statistical methods. *J. appl. Psychol.*, 1959, **43**, 395–**401.**

(Received December 9, 1960)

Journal of Applied Psychology
1961, Vol. 45, No. 6, 402-407

PROTECTION AGAINST IMPULSE-TYPE INDUSTRIAL NOISE BY UTILIZING THE ACOUSTIC REFLEX[1]

JAMES A. CHISMAN AND J. RICHARD SIMON
State University of Iowa

In recent years industrial organizations have shown a growing interest in noise control. This increased concern with the problem of noise is due, in part, to recent court decisions and legislation favorable to hearing loss compensation for workers. Though considerable progress has been made in controlling continuous noises, little has been accomplished in protecting workers from loud impulse-type noises such as those produced by drop hammers, punch presses, dynamite, etc. The distinguishing characteristic of impulse noise is its rapid rise time, i.e., only a few milliseconds elapse between the onset and the peak intensity of the noise. Impulse noise is particularly hazardous because its rise time is more rapid than the action of the intra-aural muscles which normally contract to protect the ear against loud sounds (Metz, 1951).

The purpose of the present experiment was to investigate the practicability of protecting the ear against industrial impulse noise by externally eliciting the acoustic reflex (AR) action of the intra-aural muscles prior to the presentation of the noise. The stimulus used to activate the acoustic reflex was a pure tone, hereafter referred to as an AR tone.

It has been known for some time that certain tones above approximately 70 db. sound pressure level (SPL) elicit a consensual reflex contraction of the two intra-aural (middle ear) muscles, the tensor tympani, and the stapedius. This response, known as the AR, reduces the transmission of sounds through the middle ear and acts to minimize possible cochlear damage from overstimulation by loud sounds (Wiggers, 1937).

Until recently, most investigators who have studied the intra-aural muscles have utilized animal subjects (Wever & Bray, 1937, 1942; Wever & Vernon, 1955; Wiggers, 1937). However, in recent years, research on human subjects has been stimulated by the development of new techniques such as the acoustic bridge, and by the possibility of military applications. Metz (1951) used the acoustic bridge to study the contraction pattern of the intra-aural muscles. He found that when a 1,000-cycle tone of 100 db. was presented, there was a delay of about 35 milliseconds before the intra-aural muscles began to contract and that there was an additional 100-millisecond period before the muscles reached full contraction. The time for each of these phenomena was dependent on the intensity of the activating tone, i.e., the more intense the sound, the shorter the delay and the faster the contraction. It was obvious from these results that the AR afforded very little protection against noises such as gun blasts which have rapid rise times.

Fletcher and Riopelle (1959) and Fletcher (1959) investigated the possibility of protecting the ear from the noise of machine gunfire by externally eliciting the AR action immediately before the gun was fired. In Fletcher's experiment, 13 human subjects were exposed to 100 rounds of machine gunfire, one round at a time, during a 7-minute period. Subjects were tested under three conditions: first, without an AR tone; second, with an AR tone; and third, wearing V-51R earplugs. The average SPL was 120 db., and the average peak SPL was 132 db. For the AR condition, a 1,000-cps tone of 98 db. was presented 200 milliseconds before each shot. A pure tone threshold sensitivity test was performed on each subject prior to the firing and again within 15 seconds after the firing. The differences in db. between the "before" and "after" firing threshold at each frequency was computed. This measure is known as the contemporary threshold shift (TTS). Though the

[1] This article is based on an MS thesis done by Chisman under the direction of Simon. The original thesis is on deposit at the State University of Iowa Library. The authors are indebted to Scott N. Reger for his many helpful suggestions.

TTS differs in several respects from permanent threshold shift, it is accepted by many as an indication of the permanent hearing loss which would be sustained if the subject were subjected to the same noise over a long period of time (Glorig, 1959; Jerger & Carhart, 1956).

Fletcher's results indicated that protection by the AR tone was superior to that of the earplugs up to and including 1,000 cps but markedly inferior at and above 2,000 cps. The overall mean TTS was 19.23 db. with no protection, 6.27 db. with the AR tone, and 2.50 db. with earplugs. In other words, the AR tone protected the auditory system against machine gunfire, but did not provide as much overall protection as did the earplugs.

Method

The present experiment was designed to determine whether Fletcher's notion of externally eliciting the AR action was applicable for protecting the ear against industrial impact noise. This study, however, differs from Fletcher's in several important respects:

1. Fletcher used an AR tone which had a very rapid rise time. Therefore, the AR tone itself may have produced a TTS. In the present study, the intensity of the AR activator tone was increased gradually over a period of 250 milliseconds. This gradual build-up was designed to allow the intra-aural muscles time to contract, thereby preventing the AR tone from producing a TTS.

2. Fletcher investigated only one frequency of AR tone. To determine the relative effect of frequency on the AR action, two different frequencies, 250 cps and 1,000 cps, were investigated in the present study. Previous research (Simmons, 1959) had shown that frequencies below about 2,000 cps cause the greatest reduction in transmission through the middle ear.

3. The protection afforded by earplugs was not investigated in this study since extensive research (Harris, 1957, Ch. 8, has already been conducted on the use of ear plugs for most types of industrial noise.

Apparatus

A Williams mechanical drop hammer was used to produce the impulse noise (average rise time, 1 millisecond) for this experiment. The hammer produced impacts in pairs, the two impacts in a pair being separated by approximately 370 milliseconds. Hereafter, the pair of impacts will be referred to as one impact. An Ampex 601 tape recorder was used to record 100 impacts of the hammer. To pick up the noise, a Western Electric condenser microphone was placed at the normal position for the operator. The interval between impacts was varied randomly from 2 to 6 seconds.

To reproduce the noise, the tape recorder output was fed through a 50-watt MacIntosh hi-fidelity amplifier to a 15-inch Electro-Voice extended range speaker mounted in a Karlson enclosure.[2] The impact noise was played back at 120 db. average SPL. This level was chosen so that the impact would be sufficient to produce the desired TTS and yet be below the level at which there was a danger of permanent hearing loss. The peak SPL was 139 db., and the background noise was 69 db.[3]

The AR activator tone was set at 100 db. since previous research (Metz, 1951; Wever & Vernon, 1955) suggested this to be the optimum level for AR action. To reduce the possibility of a TTS being produced by the AR tone itself, the intensity of the AR tone was increased from 70 to 100 db. over a period of 250 milliseconds.[4] The AR tone was produced by an ADC 53C-audiometer whose output was fed through a 22-watt Bell amplifier into a 12-inch GE speaker mounted in a bass reflex enclosure. The circuitry of the audiometer was revamped to obtain the desired rise time for the AR tone. The AR tone was presented approximately 400–600 milliseconds before each impact. This interval between tone and impact insured that the intra-aural muscles would be fully contracted at the time of the impact (Metz, 1951). The AR tone remained on during each impact.

Subjects

Ten college men at the State University of Iowa served as subjects. The criteria for choosing subjects were: no previous ear trouble or injuries, and no hearing loss greater than 15 db. between 250 and 4,000 cps. Seven of the 10 subjects had no losses greater than 10 db. between 4,000 and 12,000 cps. Subjects ranged in age from 22 to 31 years with a mean age of 25 years.

Procedure

The subjects served in three experimental sessions: one in which they were exposed to the recorded impact noise, one in which they were exposed to the recorded impact noise plus the 250-cps AR tone, and one in which they were exposed to the recorded

[2] The difficulties of recording and reproducing high intensity transient wave forms are well known. The evidence suggests, however, that the equipment used in this experiment was adequate for the purpose. A Sonogram spectrum analyzer and a Tektronix oscilloscope were used to compare the noise produced by the hammer with the output of the speaker. The spectra produced by the two sources were comparable, and the transient responses also compared favorably.

[3] Peak SPL was measured with a Tektronix oscilloscope in conjunction with a Western Electric condenser microphone. Average SPL and background noise were measured with a General Radio sound-level meter.

[4] A preliminary check conducted on five subjects indicated that the AR tone with the 250-millisecond rise time did not produce any TTS.

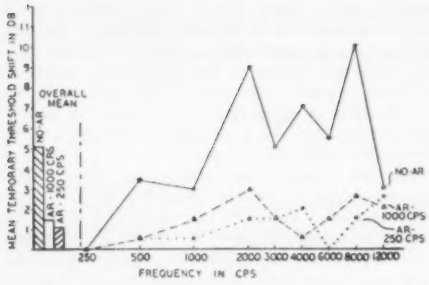


FIG. 1. Mean temporary threshold shift produced by the three experimental treatments.

impact noise plus the 1,000-cps AR tone. The order of administration of treatments was randomized.

At the start of each experimental section, an audiogram for the subject's right ear was obtained. An ADC 53C audiometer calibrated in 5-db. steps was used to determine thresholds for nine frequencies between 250 and 12,000 cps. The subject was then seated with the speaker reproducing the impact noise 1 foot from his right ear and the speaker reproducing the AR tone 1 foot from his left ear. The subject was then exposed to one of the experimental treatments in which 100 impacts were presented during a 10-minute period. Finally, a post-exposure audiogram was obtained for the subject's right ear. The audiometer test was conducted immediately following the 10-minute exposure to the noise. The order of measuring the threshold at each frequency was randomized in order to control the possibility that the TTS might wear off during the audiometer test.[5] The TTS was determined by comparing the pre-exposure audiogram with the post-exposure audiogram.

The experiment was conducted in a soundproof room. At least 24 hours elapsed between experimental sessions. A comparison of pre-exposure audiograms indicated that this period was sufficient to allow complete recovery from exposure to the noise.

RESULTS

Figure 1 shows the mean TTS produced by the three experimental treatments at each frequency measured. Note that the curves for the two AR treatments are below the curve for the No-AR treatment at all frequencies between 500 and 12,000 cps. The largest differences between the AR treatments and the No-AR treatment are at 2,000 and at 8,000 cps. The average TTS is 5.11 db. for the No-AR treatment, 1.44 db. for the 1,000-cps AR tone, and 1.11 db. for the 250-cps tone. Figure 1 also suggests that the impact noise used in this experiment produced a TTS only at frequencies above 250 cps.

Table 1 summarizes the Treatment × Frequency × Subject (T × F × S) analysis of variance used to analyze the data. Since the impact noise did not produce a TTS at 250 cps, the data for this frequency were not included in the analysis.[6] The significant interaction between Treatments and Frequency (T × F) indicates that the effect of the treatments on the TTS is not independent of the frequency at which the TTS is produced. The main effects of Treatments and Frequencies are also significant which means that the experimental treatments produced different overall effects, and exposure to the impulse noise produced a different average TTS at the different frequencies.

Inspection of Figure 1 suggests that the curves representing the two AR treatments follow a similar trend. A test of the T × F interaction for the two AR treatments alone

[5] Subsequent analysis indicated that order of determining postexposure threshold had no significant effect.

[6] The audiometer test was purposely carried out beyond the point at which a TTS might reasonably be expected to occur. This was done in order to delimit more adequately the range within which the TTS occurred. None of the treatments produced a TTS at 250 cps. On a priori grounds, therefore, the data at 250 cps were not included in the analysis of the difference between the experimental treatments. A decision to discard data based on anything other than a priori grounds would obviously modify the probability of the Type I error.

TABLE 1

SUMMARY OF ANALYSIS OF VARIANCE FOR THREE TREATMENTS

Source	*SS*	*df*	*MS*	Error Term	*F*
Treatments (T)	997.5	2	498.8	6	48.0*
Frequencies (F)	292.9	7	41.8	5	6.7*
Subjects (S)	480.8	9	53.4		
T × F	320.8	14	22.9	7	5.9*
F × S	392.5	63	6.2		
T × S	187.9	18	10.4		
T × F × S	493.8	126	3.9		
Total	3166.2	239			

* Significant at the .01 level.

(see Table 2) indicates that the trends of the two AR treatments are, in fact, parallel. A test of the Treatment main effect for the two AR treatments is also nonsignificant, indicating that the trends are not significantly displaced from one another. It is therefore concluded that the two AR treatments had the same overall effect on the subjects used in this experiment. It also follows from these results that the trend of the No-AR treatment is not parallel to the trend of either AR treatment.

Since the two AR treatments did not differ, their TTS for each frequency were averaged and compared to the TTS of the No-AR treatment at each frequency. Table 3 summarizes the results of these analyses of variance. At five of the frequencies, when the AR tone was present, the TTS was significantly less than that observed for the No-AR treatment. These frequencies are 2,000, 3,000, 4,000, 6,000, and 8,000 cps. There was no difference between the AR and No-AR treatments at three of the frequencies; 500, 1,000, and 12,000 cps. However, the difference between the treatments at 500 cps was just short of being significant.

At the end of the experiment, the subjects were asked which experimental treatment was more pleasing. The unanimous response was that the 250-cps AR tone was the most pleasing and comfortable.

TABLE 2

SUMMARY OF ANALYSIS OF VARIANCE FOR AR TREATMENTS ONLY

Source	*SS*	*df*	*MS*	Error Term	*F*
Treatments[a] (T)	5.7	1	5.7	6	0.5
Frequencies[a] (F)	64.4	7	9.2	5	1.5
Subjects[a] (S)	138.4	9	15.4		
T × F[a]	39.3	7	5.6	7	1.4
F × S[b]	392.5	63	6.2		
T × S[b]	187.9	18	10.4		
T × F × S[b]	493.8	126	3.9		

[a] The MS for these terms are based on the two AR treatments only.
[b] The MS for these error terms are the same as those in Table 1.

TABLE 3

SUMMARY OF RESULTS OF ANALYSES OF VARIANCE AT EACH FREQUENCY

Frequency	*MS*	*F*
500 cps	45.0	4.3
1,000 cps	20.0	1.9
2,000 cps	227.8	21.9**
3,000 cps	61.2	5.9*
4,000 cps	165.3	15.9**
6,000 cps	112.8	10.8**
8,000 cps	320.0	30.8**
12,000 cps	2.8	0.3

Note.—The error term used was the T × S interaction shown in Table 1.
* Significant at the .05 level.
** Significant at the .01 level.

DISCUSSION

Results of this study indicate that externally eliciting the acoustic reflex is an effective means of protecting the auditory system against an industrial impact noise. Over a wide range of frequencies, the temporary threshold shift was significantly less when an AR tone preceded each impact than when the impact alone was presented. This protection resulted despite the fact that the total acoustic stimulus was greater in the AR treatments than in the No-AR treatment.

No attempt was made in this study to compare the protection provided by an AR tone with the protection afforded by earplugs. The only study in which these two methods of protection have been compared directly indicates that earplugs afford superior protection at frequencies above 1,000 cps (Fletcher, 1959). Earplugs, however, have several serious drawbacks. First, they make normal conversation difficult. Second, they are often uncomfortable and in some cases cause infection. Third, and for the above two reasons, workers often refuse to use earplugs. Finally, in order to be effective, earplugs must be fitted to the individual. It is often difficult to attain and retain the perfect seal which is essential.

Protection by means of an AR tone has several obvious advantages over earplugs. First, AR protection operates only when necessary. Therefore, normal conversation may be carried on after the noise has terminated. Second, the AR tone, particularly the

250-cps tone, does not produce discomfort or unpleasantness. Third, the AR tone can be presented automatically. Therefore, protection does not depend on the cooperation of the worker. Finally, the AR tone can serve the additional function of acting as a pre-impact warning device for workers in the area.

The results of this limited investigation when viewed in the light of the many disadvantages of earplugs suggest that activating the acoustic reflex may be a practical and effective way of protecting industrial operators from impact noises. With most industrial machines of the impact type, the operator is confined to a limited area. It would not be difficult to install a tone generator near the operator's ear and, by means of an electronic timer integrated with the machine, to control the time relations between the AR tone and the impact. The timer could also be employed to cut off the AR tone an instant before each impact, thus reducing the amount of energy at the moment of impact. The operator would still be afforded maximum protection since the intra-aural muscles remain fully contracted for about 250 milliseconds after the AR tone has ceased (Metz, 1951).

One final point deserves some comment. The overall TTS produced in this study was substantially less than that observed in Fletcher's study although the average noise levels reported in the two studies were similar. There are several possible reasons for the differing results. First, Fletcher points out that, because of the instruments he used, his estimates of SPL were conservative (Fletcher & Riopelle, 1959). The instruments used in the present experiment were more appropriate for measuring noise with a rapid rise time. Second, the 139-db. peak SPL reported in the present experiment is not comparable to the 132-db. *average* peak SPL reported by Fletcher. In a later experiment which employed Fletcher's setup, (Hilding, 1960), the *peak* SPL was estimated to be 155–160 db. which is considerably higher than the 139-db. peak in the present study. Third, the noise spectra used in the two studies were considerably different. Fletcher's noise spectrum was at the high end of the audio range, and the noise spectrum in the present study was at the low to middle end. Fourth, Fletcher presented more impulses per unit of time, i.e., 100 impulses in 7 minutes versus the 100 impulses in 10 minutes used in the present study.

Summary

This experiment was designed to investigate the practicability of externally eliciting the acoustic reflex to protect the ear from industrial impulse noise. Ten subjects were exposed to 100 impacts of the recorded noise of a mechanical drop hammer during a 10-minute period. The noise was presented at a 120-db. average SPL. Subjects listened to the same noise under three conditions. In Condition 1, the impacts were presented alone, i.e., with no AR tone presented. In Condition 2, a 250-cps, 100-db. tone was presented between 400 and 600 milliseconds before each impact. Condition 3 was the same as Condition 2 except that a 1,000-cps tone was used. Preexposure audiograms were compared with postexposure audiograms to determine the temporary threshold shift (TTS) under each condition.

Analysis of the data indicated a significant difference in TTS between the No-AR condition and the two AR conditions at the following frequencies: 2,000, 3,000, 4,000, 6,000, and 8,000 cps. No differences were observed at 250, 500, 1,000, and 12,000 cps. In terms of overall protection, the 250-cps AR tone did not differ from the 1,000-cps AR tone although subjects found the 250-cps tone to be more pleasing.

Results are interpreted to suggest that elicting the acoustic reflex prior to an impact may be an effective means of protecting the ear against industrial impulse noises. The advantages of AR protection over protection provided by earplugs are discussed.

References

Fletcher, J. L. Comparison of attenuation characteristics of the acoustic reflex and the V-51R earplug. *USA Med. Res. Lab. Rep.*, 1959, No. 397.

Fletcher, J. L., & Riopelle, A. J. The protective effect of acoustic reflex for impulsive noises. *USA Med. Res. Lab. Rep.*, 1959, No. 396.

Glorig, A. Current research in industrial noise. *Noise Control*, 1959, 5(1), 32–35, 74.

Harris, C. M. (Ed.) *Handbook of noise control*. New York: McGraw-Hill, 1957.

HILDING, D. A. The intratympanic muscle reflex as a protective mechanism against loud impulsive noise. *Ann. Otol. Rhinol. Laryngol.*, 1960, **69**, 51–60.

JERGER, J. F., & CARHART, R. Temporary threshold shift as an index of noise susceptibility. *J. Acoust. Soc. Amer.*, 1956, **28**, 611–613.

METZ, O. Studies on the contraction of the tympanic muscles as indicated by changes in the impedance of the ear. *Acta otolaryngol., Stockh.*, 1951, **39**, 397–405.

SIMMONS, F. B. Middle ear muscle activity at moderate sound levels. *Ann. Otol. Rhinol. Laryngol.*, 1959, **68**, 1126–1143.

WEVER, E. G., & BRAY, C. W. The tensor tympani muscle and its relation to sound conduction. *Ann. Otol. Rhinol. Laryngol.*, 1937, **46**, 947–961.

WEVER, E. G., & BRAY, C. W. The stapedius muscle in relation to sound conduction. *J. exp. Psychol.*, 1942, **31**, 35–43.

WEVER, E. G., & VERNON, J. A. The effects of the tympanic muscle reflexes upon sound transmission. *Acta otolaryngol., Stockh.*, 1955, **45**, 433–439.

WIGGERS, H. C. The functions of the intra-aural muscles. *Amer. J. Physiol.*, 1937, **120**, 771–780.

(Received December 16, 1960)

Journal of Applied Psychology
1961, Vol. 45, No. 6, 408–414

A FACTOR ANALYSIS OF THE CALIFORNIA PSYCHOLOGICAL INVENTORY

JOHN O. CRITES, HAROLD P. BECHTOLDT, LEONARD D. GOODSTEIN, AND ALFRED B. HEILBRUN, JR.

University of Iowa

Although developed over a period of years by Gough (1957) and based in part upon research with the Minnesota Multiphasic Personality Inventory, the California Psychological Inventory (CPI) is a relatively unstudied newcomer to the area of personality measurement. A multidimensional inventory of normal personality characteristics, the CPI consists of 18 separate scales grouped into four areas of "adjustment and development." The first area, which describes aspects of social functioning, includes measures of Dominance (Do), Capacity for Status (Cs), Sociability (Sy), Social Presence (Sp), Self-Acceptance (Sa), and Sense of Well-Being (Wb). The second area, which concerns emotional and social development, is comprised of scales on Responsibility (Re), Socialization (So), Self-Control (Sc), Tolerance (To), Good Impression (Gi), and Communality (Cm). The third area, which relates to actual and potential achievement in educational and occupational pursuits, contains measures of Achievement via Conformance (Ac), Achievement via Independence (Ai), and Intellectual Efficiency (Ie). The final area, which assesses intellectual and interest modes, has measures of Psychological-Mindedness (Py), Flexibility (Fx), and Femininity (Fe). The various CPI scales were grouped into these four areas in order to "emphasize some of the psychological and psychometric clustering which exist among them" (Gough, 1957, p. 7). The groupings were not based upon factor analysis, however, but rather upon inspection of intercorrelations among scales.

PROBLEM

In this study factor analysis of the CPI scale intercorrelations was used to accomplish a number of purposes. The major intent of the investigation was to test Gough's subjective groupings of CPI scales against the more objective clusters derived from a factor analysis. An additional objective was to reduce the 18 scales of the CPI to a more economical set which would not only result in more efficient testing but which would facilitate pattern analysis in the interpretation of score configurations on the CPI profile. A final purpose was to evaluate the extent to which the CPI assesses different facets of normal personality functioning.

PROCEDURE

Subjects

A sample of 372 subjects, stratified along two dimensions, sex and client status, was available for the factor analysis. With respect to the sex variable, males and females were treated separately for two reasons: first, raw scores and hence norms for the two groups were different (Gough, 1957); and, second, it was thought that the scale interrelationships for the sexes might differ, since different social role expectations for men and women probably result in unique configuration of personality characteristics. With regard to client status, client and nonclient groups were selected for independent analysis, because a previous study indicated that they differ in both kinds and degrees of adjustment problems (Goodstein, Crites, Heilbrun, & Rempel, 1961). One group of clients, which was designated "personal" (P), included all individuals who upon intake at a university counseling center checked the following category on an application card: "Would like to discuss my feelings about myself and others; some personal problems which are bothering me." The second group of clients, which was termed "vocational-educational" (VE), consisted of persons who at intake endorsed either or both of these statements: "Would like to discuss my abilities, interests, aptitudes; my occupational plans for the future," and "Would like to discuss my courses, grades, classes, study techniques; my progress here at the University." Following initial application for counseling, but preceding interviews with a counselor, the CPI was routinely administered to all clients. The nonclients, or controls (C), were members of a sophomore and junior level psychology course, who were tested with the CPI during a class period. The total sample of 372 was stratified, then, according to the two sexes and three categories of client-nonclient status to yield six groups of 62 subjects each for the factor analysis.

Design

Scores on the 18 CPI scales for all subjects in the various groups were first punched on cards for use with the IBM 650 computer. The next step in the analysis was to test the homogeneity of the variance-covariance matrices for the six samples (Anderson, 1958). If these were relatively homogeneous, then the groups could be combined for the factor analysis proper. The findings indicated, however, that the assumption of homogeneity was untenable. To locate the sources of heterogeneity, the variance-covariance matrices were compared for the six samples stratified by sex and client-nonclient status. In these comparisons, there were no differences between sexes which were significant at the .01 level, although one was significant in the C group at the .05 level. There were significant differences at the .01 level, however, between the P, VE, and C groups in all possible combinations. It was decided, therefore, to combine the male and female data (average within-group correlations) for the client-nonclient groups and to perform the factor analyses only on the latter. The remaining steps in the procedure involved: (*a*) tests of independence, using the determinants of the population estimates of the variance-covariance and correlational matrices (Anderson, 1958); (*b*) computation of factors for the three client-nonclient groups, following the complete centroid method (Thurstone, 1947); (*c*) tests for number of significant factors based on Lawley's maximum-likelihood procedure (Thomson, 1951); and (*d*) analytical and graphical rotation of the factors to an oblique simple structure (Thurstone, 1947).

In the selection of reference tests for each of five factors, two criteria were used. A set of CPI scales which had the largest loadings on the factors was first identified, and then selections from the set were made, one for each factor, to minimize the number of overlapping items in the various scales. A sixth reference test was selected to represent a dimension not defined by the five-factor scales. Thus, a scale chosen to represent a factor or dimension meets two requirements: (*a*) it has a high, although not necessarily the highest, loading on a factor; and (*b*) it has the fewest number of items in common with scales which represent other factors. The latter criterion was necessary because the empirical approach used by Gough in the construction of most of the CPI scales resulted in considerable item overlap, a variable which becomes confounded with the statistical interrelationships of the scales when they are correlated. Unless this lack of experimental independence is eliminated, it is difficult to analyze the common variance among the scales.

To evaluate the adequacy of the CPI reference scales which were finally selected, multiple correlations between the selected set and the remaining scales were computed. These multiple *R*s indicated the extent to which the selected scales reproduced performance on the unselected scales and hence were representative of different scale groups or clusters. Furthermore, the R^2s provided an estimate of the amount of the total variance accounted for by the reference scales.

Results

The tests of independence for each of the correlation matrices of the P, VE, and C groups, as presented in Table 1, were significant. All of the determinants for the groups exceeded the expected value of .267 ($SE = .0402$) at the .01 level. These findings, which indicated significant degrees of dependence in the matrices, supported factorial analyses of the CPI scale intercorrelations for the three client-nonclient groups.

For each of the three groups, the first six of eight centroid factors were significant according to Lawley's (1951) test after iteration through two cycles by Bargmann's (1957) procedure. With 60 *df*, the χ^2 values of 216.70, 168.77, and 140.92 for the P, VE, and C groups, respectively, exceeded the .001 level. Since the residuals were less than $\pm .10$, it appeared that the six factors accounted for most of the common variance of the CPI scales and that oblique rotations of five or six factors were justifiable. The first analytical approximation to a simple structure was poor, however, probably because many of the scales were not experimentally independent and hence formed an unusual number of doublets, but successive graphical rotations eventually yielded solutions which were more adequate and which were quite uniform for the three groups. The best solution was based upon the five factors listed in Table 2, although the communalities (h^2) for the six centroid factors indicated that unaccounted for residual variance remained. Note that there are some variations in the factor loadings across groups, but that these are considerably less marked than the consistencies.

From the factor matrices five scales, which had the highest factor loadings *and* fewest overlappings items, were selected as a reduced set of CPI variables for a multiple regression analysis of their relationships to the remaining scales. The selected set of scales included the following: Dominance (Do), Good Impression (Gi), Communality (Cm), Flexibility (Fx), and Femininity (Fe). With these five scales as the predictor variables and the other 13 scales as the criterion variables, beta

TABLE 1

INTERCORRELATIONS OF THE CALIFORNIA PSYCHOLOGICAL INVENTORY SCALES FOR THE PERSONAL (P), VOCATIONAL-EDUCATIONAL (VE), AND CONTROL (C) GROUPS

(For each group $df = 122$)

Scale		Do	Cs	Sy	Sp	Sa	Wb	Re	So	Sc	To	Gi	Cm	Ac	Ai	Ie	Py	Fx	Fe	Mean	*SD*
	P																			47.50	12.43
Do	VE																			50.16	12.44
	C																			54.96	11.46
	P	55																		50.16	10.51
Cs	VE	55																		50.62	10.41
	C	64																		53.90	11.03
	P	71	64																	44.90	11.11
Sy	VE	70	63																	50.01	10.36
	C	64	63																	55.25	10.41
	P	61	64	69																49.32	12.56
Sp	VE	56	64	73																51.09	11.30
	C	52	62	64																57.38	11.49
	P	71	55	65	67															54.40	11.12
Sa	VE	68	59	75	67															55.58	11.08
	C	66	48	53	63															59.53	10.75
	P	31	36	36	27	12														40.24	14.49
Wb	VE	30	31	23	16	11														45.94	12.13
	C	22	37	41	29	03														49.76	10.18
	P	14	24	18	13	−12	53													45.96	10.62
Re	VE	28	25	09	−04	−03	50													49.66	8.23
	C	27	30	32	04	−07	54													49.59	9.83
	P	09	10	13	11	−14	61	61												40.28	11.25
So	VE	15	11	03	−02	−03	56	47												48.16	10.43
	C	27	27	32	00	02	50	64												50.52	9.89
	P	−07	07	03	−19	−30	63	66	67											40.22	12.41
Sc	VE	01	11	−17	−17	−28	68	48	59											44.54	10.66
	C	−06	−06	13	−16	−33	62	60	54											44.26	10.11

TABLE 1—Continued

Scale		Do	Cs	Sy;	Sp	Sa	Wb	Re	So	Sc	To	Gi	Cm	Ac	Ai	Ie	Py	Fx	Fe	Mean	*SD*
	P	24	50	35	25	08	72	64	48	64										45.83	12.62
To	VE	38	58	33	37	21	69	53	37	56										48.10	9.90
	C	20	20	37	30	06	64	54	37	53										51.36	11.29
	P	16	24	26	01	−08	56	59	53	80	58									40.86	10.65
Gi	VE	26	36	20	11	−03	59	44	38	77	60									43.84	9.80
	C	07	07	29	−06	−20	62	51	49	83	47									44.84	10.58
	P	19	05	20	−03	02	30	39	39	29	32	15								47.35	9.79
Cm	VE	01	−04	13	03	15	11	06	32	04	03	−05								51.88	8.18
	C	09	09	11	06	01	28	37	41	23	14	13								51.87	9.95
	P	30	32	35	11	09	64	66	63	69	60	70	33							41.78	11.24
Ac	VE	42	35	29	18	23	71	47	60	60	60	57	24							45.60	10.81
	C	52	52	50	24	26	52	54	62	49	53	50	21							48.34	10.69
	P	18	54	29	33	19	45	47	22	41	71	42	10	55						53.60	11.00
Ai	VE	27	48	16	22	17	51	36	17	40	70	41	04	55						50.56	10.02
	C	29	29	20	31	19	48	37	25	40	64	27	11	51						54.32	10.20
	P	36	63	51	44	28	66	54	43	45	76	44	28	61	69					45.00	12.69
Ie	VE	54	60	51	68	43	66	40	47	41	74	46	11	67	66					45.33	12.72
	C	39	39	53	51	32	62	51	43	44	64	38	17	60	68					52.39	11.52
	P	37	50	36	43	25	52	34	22	41	55	38	11	54	62	60				45.86	12.60
Py	VE	48	45	37	59	32	54	32	30	36	60	42	−05	56	51	67				43.50	12.12
	C	43	43	39	41	21	48	35	28	34	42	27	08	43	50	54				51.14	9.83
	P	05	33	12	48	23	11	−09	−13	−14	22	−07	−28	−03	44	33	21			60.06	11.26
Fx	VE	−05	22	08	75	04	04	−11	−25	−12	28	−03	−09	−01	41	23	23			54.40	10.63
	C	−02	−02	01	30	08	−01	−22	−36	−11	14	−13	−16	−12	38	18	19			54.90	10.90
	P	−11	−15	−26	−26	00	−10	14	14	04	−10	−03	−01	11	−01	−08	−17	01		51.64	10.60
Fe	VE	−16	−22	−30	67	16	−11	16	10	07	−12	−01	−09	−02	−10	−30	−21	−27		50.96	10.30
	C	−06	−06	−17	−30	−06	−07	22	12	13	−02	03	10	02	07	−01	−15	−09		48.10	10.21

Note.—Decimal points omitted.

weights and *R*s were obtained for the P, VE, and C groups combined ($df = 366$) since the variance-covariance matrices for the predictors were homogeneous. The residual intercorrelations among the criterion variables in this analysis revealed that Intellectual Efficiency (Ie) was highly related to each of them and moderately related to the five factor scales. Thus, Intellectual Efficiency was included in the selected set of CPI scales as a variable which may account for part of the obliqueness among the factors as well as the sixth factor variance. Table 3 lists the beta weights and R^2s for the prediction of each of the 12 unselected scales from the combined 6 reference scales. The R^2s range from .430 to .734 and average .542. The sum of the specific and common variance predicted by the selected set accounts for approximately 66% of the total variance in the 18 CPI scales.

TABLE 2

FINAL ROTATED OBLIQUE FACTOR LOADINGS OF THE CALIFORNIA PSYCHOLOGICAL INVENTORY (CPI) SCALES FOR THE PERSONAL (P), VOCATIONAL-EDUCATIONAL (VE), AND CONTROL (C) GROUPS

CPI Scales	Groups	Factors A	B	C	D	E	h^2[a]	CPI Scales	Groups	Factors A	B	C	D	E	h^2[a]
	P	08	74	−09	−09	−02	78		P	07	−06	33	51	22	85
Do	VE	05	72	19	08	−09	79	To	VE	35	27	04	38	06	82
	C	15	76	−10	−08	−04	80		C	11	−05	26	48	18	64
	P	04	37	−02	39	13	67		P	61	22	−15	−04	06	73
Cs	VE	23	66	−06	14	09	71	Gi	VE	69	25	−09	−06	03	74
	C	06	48	04	27	15	72		C	67	−10	−12	−11	08	81
	P	04	62	01	00	20	80		P	−12	06	39	−11	06	33
Sy	VE	15	77	20	−11	22	82	Cm	VE	−08	−12	58	−03	08	41
	C	16	50	00	03	35	76		C	−09	−11	43	06	06	30
	P	02	44	−18	26	36	86		P	45	33	01	04	−09	79
Sp	VE	09	66	05	03	36	75	Ac	VE	22	10	48	29	00	76
	C	−22	29	08	40	51	84		C	38	44	07	04	−07	70
	P	−05	64	−16	08	−08	75		P	19	02	−01	70	−01	84
Sa	VE	−13	69	24	04	06	76	Ai	VE	−02	04	03	70	−02	81
	C	−14	57	−05	07	06	64		C	03	07	14	65	−10	83
	P	21	10	22	15	35	76		P	02	10	26	51	21	81
Wb	VE	38	07	30	23	10	71	Ie	VE	20	33	28	38	24	83
	C	29	−04	21	28	36	72		C	06	17	24	50	16	75
	P	12	05	34	17	−16	74		P	33	19	−14	36	21	60
Re	VE	22	11	14	20	−31	48	Py	VE	14	26	15	38	17	66
	C	20	04	39	11	−04	66		C	26	24	−08	24	26	51
	P	17	05	32	−07	02	65		P	−07	−17	−07	59	20	51
So	VE	27	−06	53	01	−10	64	Fx	VE	−13	−07	−20	43	35	52
	C	31	17	32	−12	−05	72		C	−21	−08	−12	55	07	51
	P	50	−04	07	−01	11	91		P	01	05	00	02	−51	35
Sc	VE	65	−15	06	03	01	92	Fe	VE	−09	−08	−06	−03	−58	39
	C	58	−10	04	06	−01	90		C	−11	−05	24	04	−46	30

Note.—Decimal points omitted.

[a] Communalities are for six centroid factors.

TABLE 3

BETA WEIGHTS AND MULTIPLE *R*s FOR THE REFERENCE AND UNSELECTED CALIFORNIA PSYCHOLOGICAL INVENTORY SCALES BASED UPON THE COMBINED GROUPS ($df = 366$)

Scale	Do	Gi	Cm	Fx	Fe	Ie	R^2
Cs	404**	068	−050	149**	−060	378**	533
Sy	558**	065	049	016	−149**	217**	554
Sp	452**	−155**	−021	259**	−171**	281**	533
Sa	660**	−279**	−018	057	031	171**	542
Wb	001	370**	109**	−017	−035	467**	545
Re	011	334**	171**	−152**	206**	370**	463
So	−038	294**	245**	−239**	131**	329**	430
Sc	−272**	716**	093**	−110**	073*	263**	734
To	−026	318**	073*	108**	003	548**	593
Ac	168**	376**	119**	−090*	102**	404**	582
Ai	−008	154**	028	303**	091*	548**	553
Py	213**	154**	−047	096*	−089*	422**	436

Note.—Decimal points omitted.
* .05 level of significance.
** .01 level of significance.

DISCUSSION

The results of the factor analyses both agree and disagree with Gough's clusters of scales for the CPI. With one exception, the Dominance scale clearly represents the Class I scales which Gough terms "Measures of Poise, Ascendancy, and Self-Assurance" and which include the Capacity for Status, Sociability, Social Presence, and Self-Acceptance scales. The exception is the Well-Being scale, which Gough classified with these variables, but which has higher empirical relationships with the Responsibility, Socialization, Self-Control, and Tolerance scales, group of measures best defined by the Good Impression scale. Closely related to this group, but yet distinct from it, the Communality scale probably belongs with these Class II "Measures of Socialization, Maturity, and Responsibility." In Class III, "Measures of Achievement Potential and Intellectual Efficiency," the Intellectual Efficiency scale correlates highly with Achievement via Conformance and Achievement via Independence, but it also relates well to most of the other scales in the CPI. The final group of scales, the Class IV "Measures of Intellectual and Interest Modes," breaks down into quite specific factors, as assessed by the Flexibility and Femininity scales, which have only negligible relationships to each other and to the other measures. Thus, the findings support two of the four scale clusters or classes which Gough proposes but offer no evidence for the existence of the other two.

The six scales which emerge from the factor and multiple regression analyses provide an economical but comprehensive coverage and evaluation of the major aspects of the normal personality. Two of the scales, Dominance and Good Impression, seem to measure markedly different *modes of adjustment* or ways of coping with the environment and others. The Dominance scale reflects an inclination to master and control people and things (Horney, 1945) and a tendency to direct and manipulate affairs: adjustment comes from changes in external reality rather than in the self. In contrast, the Good Impression scale defines an attitude of compliance, cooperation, and conformity (Riesman, 1950): adjustment derives from adaptations of the self to reality. Though distinctive, these two adjustment modes appear equally effective, as indicated by the positive relationships of the Dominance and Good Impression scales to the Intellectual Efficiency scale which apparently assesses a *general level of adjustment* variable and which corresponds to the over-all elevation of the CPI profile. The Femininity and Flexibility scales supplement these measures of adjust-

ment patterns and status with scores on two highly specific attitudinal and trait dimensions. Finally, the Communality scale, despite its relationship to the Good Impression group, contributes unique variance to the reference scale set and probably measures the tendency to make "modal" responses to the CPI, as stated in the manual (Gough, 1957). The scales yielded by the factor and correlational analyses, then, survey most of the variables important in the description and definition of the normal personality: adjustment modes, adjustment level, attitudes and traits, and test-taking set.

That the various statistical analyses indicate that a reduced set of six scales accounts for about two-thirds of the total variance in the CPI measures strongly suggests the use of the reference scales for more economical testing and more meaningful score interpretations. Not only are these scales relatively independent statistically, but they are almost wholly unrelated experimentally. With the exception of the Intellectual Efficiency and Dominance scales, which have two items in common, there is no item overlap among the six scales. As a result, the individual scales are more sharply defined, and pattern analyses of scores on combinations of scales are more clearly formulated. Since the mean values of the three groups tend to differ on these selected scales, further discriminant function analyses may be of value. The ultimate meaning and usefulness of the reference scales depends, of course, upon their relationships to other, external variables. Research currently in progress deals with the problem of validating the reference scales and constitutes the next phase in this series of studies on the CPI and the measurement of the normal personality.

Summary

The purposes of this study were to test with factor analytic methods the rational groupings of the 18 scales of the California Psychological Inventory (CPI) into four areas or classes and to determine the number of independent variables measured by the scales. The sample consisted of 372 male and female subjects classified into three groups: personal adjustment clients, vocational-educational clients from a university counseling center, and nonclient controls from an undergraduate psychology course. The factor analyses, which were performed on the intercorrelations of the CPI scales for each group separately ($df = 122$), yielded a reduced set of six reference scales which exhibited homogeneous correlation matrices for all groups. These scales, which included measures of Dominance, Good Impression, Intellectual Efficiency, Flexibility, Femininity, and Communality, represented group, general, and specific factors which corresponded to adjustment modes, adjustment level, personal traits, and test-taking attitude. The findings supported some of the original scale clusters but not others and indicated that the six reference scales predict most of the reliable variance in the other CPI measures. The conclusion drawn from the results was that the reference scales provide more concise and less ambiguous definitions of the variables assessed by the CPI and have considerable promise as measures of the normal personality.

References

Anderson, T. W. *Introduction to multivariate statistical analysis.* New York: Wiley, 1958.

Bargmann, R. A study of independence and dependence normal analysis. University of North Carolina, Institute of Statistics, 1957, Ser. No. 186. (Mimeo)

Goodstein, L. D., Crites, J. O., Heilbrun, A. B., Jr., & Rempel, P. P. The use of the California Psychological Inventory in a university counseling service. *J. counsel. Psychol.*, 1961, **8**, 147–153.

Gough, H. G. *California Psychological Inventory manual.* Palo Alto, Calif.: Consulting Psychologists, 1957.

Horney, Karen. *Neurosis and human growth.* New York: Norton, 1950.

Lawley, D. N. The maximum likelihood method of estimating factor loadings. In G. Thomson (Ed.), *The factorial analysis of human ability.* (4th ed.) New York: Houghton Mifflin, 1951.

Riesman, D. *The lonely crowd.* New Haven: Yale Univer. Press, 1950.

Thurstone, L. L. *Multiple-factor analysis.* Chicago: Univer. Chicago Press, 1947.

(Received December 23, 1960)

Journal of Applied Psychology
1961, Vol. 45, No. 6, 415–419

CHANCE VERSUS NONCHANCE SCORES ON THE SVIB[1]

RICHARD R. STEPHENSON[2]

University of Iowa

Scores on the Strong Vocational Interest Blank (SVIB) are usually reported on either of two specially devised profile forms, the Hankes Report Form or the IBM Report Form. These special profiles (Strong, 1959)

> indicate which scores are significant and which are not, for the . . . [profiles are] printed so that all *nonsignificant* [italics added] scores must be plotted upon a gray background whereas significant scores will fall upon a white background (p. 7).

The gray background indicates the "chance" range. The chance ranges were

> determined by marking 40 blanks on the basis of dice throwing and the shaded areas give the middle 68 per cent of the distribution of such chance scores for each scale. Scores which fall in the shaded areas may be interpreted as easily obtainable by chance and therefore indeterminate for the particular scale in question (p. 7).

While the profile forms may be shaded on the basis of the above method, the chance means may be more accurately located by taking the standard score equivalents of one-third of the algebraic sums of the scoring weights for each possible response to each item for each occupational scale (Strong, 1959, p. 34). Mathematically exact standard deviations for the distribution of chance scores have been published by Lyerly (1957). Thus, for most occupational scales, it is possible to precisely delimit the chance range (mean, plus and minus one standard deviation) in standard score units for each occupational scale, which makes more relevant the question of their significance.

If, as Strong indicates, scores which fall in the chance range are nonsignificant, then it becomes important to assess how much of one's data is "nonsignificant" for a given counseling client and to compare this result with relevant normative data. It becomes even more important to have some estimate of the amount of change to be anticipated in the event that retest is decided upon in the hopes of obtaining a "significant" finding. Finally, and perhaps most important, if a given SVIB profile is composed of two different "kinds" of data, chance and nonchance, then it may be necessary to re-evaluate all of the research findings on SVIB stability. The implication in all of these questions is that if a score falls in the chance zone because of "chance" factors (factors that are unknown or random), then, on retest, such scores might shift to any other location on the given occupational scale. Nonchance scores, on the other hand, ought to vary on retest only within the limits of the standard error of measurement of the scale involved.

METHOD

While the above considerations form the general background for the present study, the specific impetus was found in a study (Stephenson, 1961) done on high ability, male, arts college freshmen. To obtain SVIB letter grade normative data for this group, such letter grades were tabulated under the Categories A, B+, B, B− and C+, C, and Chance. The B− and C+ letter grades were combined into a single category as a result of Strong's (1943) observation that scores falling in these categories are "neutral and . . . can very largely be ignored" (p. 435). The Chance category was used to denote a score falling in the chance area. The intent was to exclude these nonsignificant, chance, scores from the analysis. However, there were more scores in the Chance category than in any other single category! These data provided the nucleus of the normative data.

Next, SVIB data, made available by Powers (1956) and by King (1957), were inspected. Powers' sample consisted of 109 males tested in 1931 and again in 1941. Most of these men were in nonprofessional noncollege work, and all had a history of unemployment sometime during this 10-year span.

[1] Based in part upon a paper presented at the 1961 American Personnel and Guidance Association convention in Denver, Colorado.

[2] The writer wishes to express his gratitude to Leslie A. King of the General College of the University of Minnesota and Mabel K. Powers of the Senior Division, College of Science, Literature, and the Arts, of the University of Minnesota, for making available their data without which the present research would not have been possible; and to Richard L. Walker, Research Assistant, for his assistance in the tedious task of retabulating the above data.

King's sample consisted of 242, generally low-ability college freshmen, who were tested in 1954 and again, approximately 9 months later, in 1955. Results of the first testing of both Powers' and King's samples were tabulated separately according to the SVIB letter grade categories described for Stephenson. These data provided additional chance score normative data.

Finally, 100 SVIB answer sheets were scored by use of the Rand Corporation's (1955) table of random numbers, a million digit series. This series is large enough that no digit series was used more than once. The first 9 digits were used to answer the Like, Indifferent, Dislike questions, and all 10 digits the Most Liked-Least Liked response sections. These data complete the normative study.

To assess the relative stability of chance versus nonchance scores, test-retest data were necessary. As indicated, these were secured from Powers and King. Each test score (initial testing) for each subject in both samples was tabulated on each of 44 occupational scales of the SVIB, using the following procedure: Each score was first dichotomized as being in the chance range, using algebraically derived means, plus and minus mathematically exact standard deviations, or out of the chance range. This standard score was tabulated. The retest tabulations, then, were simply in terms of SVIB standard scores, and were computed for each subject and for each scale. This procedure provides an answer to the question: Do scores that have fallen in chance areas upon first testing differ more at retest than do scores that have originally fallen outside the chance area?

The rationale for the above question is quite simple: If chance is chance, or indifference, or anything but measurement, then there is no prediction as to where such scores will fall upon retest. On the other hand, for scores that originally fall outside the chance area, retest deviations should reflect only predictable errors of measurement. Nonchance scores, then, provide the criterion against which to answer the above research question. The absolute deviation was chosen as the appropriate measure since, under the assumption, both chance and nonchance scores will be as likely to vary in one direction as in the other, and an algebraic measure would, therefore, average to zero.

Findings

For greater ease in making visual comparison, SVIB letter grade normative data for all normative samples have been converted to percentages and are presented in Table 1. As can be seen in Table 1, the best guess of a subject's score for any scale is chance. Even considering that Stephenson's sample consists of high ability college males, that King's consists primarily of low ability college males, and that Powers' consists primarily of noncollege males, it seems that chance as a modal score is a common phenomenon.

A common distributions chi square test of the three male samples in Table 1 resulted in a chi square of 173.104, which, with 10 *df*, is significant far beyond the .001 level. Thus, it appears that, in groups that differ widely in scholastic aptitude, significant differences are to be expected in relative proportions found within each SVIB letter grade category, including the category Chance. However, the rank-difference correlation (rho) between King's sample and Powers' sample is 1.0 and between either of these and Stephenson is .94.

Table 2 summarizes the findings when the retest absolute deviations are compared over SVIB occupational scales, and may be interpreted as indicating that, while average scale deviations are significantly larger after 10

TABLE 1

Percentages of Scores in the Chance Zone or in One of the Various Letter Grade Ratings on 44-Scale SVIB Profiles for Three Different Samples of Male Persons

SVIB letter grade	Noncollege males[a] (N = 109)	College low ability males[b] (N = 242)	College high ability males[c] (N = 239)	Blanks scored with table of random numbers (N = 100)
A	11.5	11.7	10.2	0.5
B+	8.5	9.3	9.4	3.8
B	10.5	10.9	12.5	7.0
B-C+	14.0	13.9	18.6	13.6
C	26.4	23.7	28.3	14.3
Chance	29.1	30.5	21.0	60.8

Note.—The base for each percentage is the sample size × 44.
[a] From Powers (1956) first testing.
[b] From King (1957) first testing.
[c] From Stephenson (1961).

TABLE 2

SUMMARY OF FINDINGS WHEN AVERAGE ABSOLUTE DEVIATIONS ARE COMPARED OVER SVIB OCCUPATIONAL SCALES

Variable	Low ability college subjects[a] (N = 242)		Noncollege subjects[b] (N = 109)	
	In chance	Not-in chance	In chance	Not-in chance
N of scales	44	44	44	44
N of differences	3251	7397	1394	3402
Most variable scale's average absolute deviation	8.84[c]	7.79[d]	13.08[e]	9.73[e]
Least variable scale's average absolute deviation	4.17[f]	3.95[g]	4.88[h]	4.98[i]
Total scales' average absolute deviations	252.46	266.32	333.77	323.02
Average absolute deviation per scale	5.73	6.05	7.58	7.34
Standard deviation of scale averages	1.06	.88	1.86	1.21
t	1.518 $p > .05$		.734 $p > .05$	
Analysis of variance	$SS_b = 111.7163$ $df = 3$ $SS_w = 293.0395$ $df = 172$		$F = 21.8576$ $p < .01$	

[a] Data from King (1957).
[b] Data from Powers (1956).
[c] Social science high school teacher.
[d] Personnel director.
[e] Industrial arts teacher.
[f] Banker.
[g] Real estate salesman.
[h] Banker.
[i] Lawyer.

years than after 9 months, for a given period of time there is no significant difference between scores that fall in chance on a particular scale and scores that fall out of chance on that scale.

It should be noted that Table 2 summarizes findings based on four distributions of 44 scales each. Each scale average is based upon the N that happened to fall in or out of the chance range on that particular scale. These Ns vary from 7 (Powers: in-chance, Carpenter) to 206 (King: not-in-chance, Author-Journalist, and Forest Service Man). Further, the occupational scales themselves have differing reliabilities. Thus, there may be significant individual differences in chance and nonchance variability that are masked by collapsing the data over scales. The question that now must be answered is this: Are there significant differences between the in-chance variability and the not-in-chance variability of these individuals *over* these 44 occupational scales? The data to answer this question are presented in summary fashion as Table 3.

Table 3 may be interpreted as indicating that individuals are more variable after 10 years than after 9 months, but, after this long, are no more variable on their originally in-chance scales than on their originally not-in-chance scales, which is the same conclusion reached when the data were collapsed over scales.

Note, however, that in Table 3 there is a significant mean difference between the in-chance and the not-in-chance variability of the two groups in King's sample. In explaining this significant difference, at least two

TABLE 3

SUMMARY OF FINDINGS WHEN AVERAGE ABSOLUTE DEVIATIONS ARE COMPARED OVER PERSONS

Variable	Low ability college subjects[a]		Noncollege subjects[b]	
	In chance	Not-in chance	In chance	Not-in chance
N of persons	242	242	109	109
N of scales	44	44	44	44
Absolute deviations of most variable subject	16.5	20.5	16.5	18.5
Absolute deviations of least variable subject	1.5	2.5	2.5	3.5
Total group's absolute deviations	1364.0	1475.0	799.5	810.5
Average deviation per subject	5.64	6.10	7.33	7.44
Standard deviation of distribution of subjects	2.02	2.20	2.63	2.84
z	2.39 $.02 > p > .01$		.30 $p = .77$	
Analysis of variance	$SS_b = 373.1 \quad df = 3$ $SS_w = 3794.4 \quad df = 698$		$F = 22.866 \quad p < .01$	

[a] Data from King (1957).
[b] Data from Powers (1956).

thoughts come immediately to mind: First is that the absolute magnitude of the significant difference is less than one standard score point. This small size questions the practical significance of the statistical finding. Second, the direction of the difference is opposite to that predicted by the hypothesis under investigation: that in-chance scores are the more variable. Thus, though new hypotheses may be generated by this finding, the general conclusions of the present investigation, that in-chance scores are *not* more variable than not-in-chance scores, cannot be altered. (The negligible differences in means between Tables 2 and 3 are the result of grouping errors in Table 3 and rounding errors in both tables.)

DISCUSSION

The most obvious fact to emerge from analysis of these data is the evidence, perhaps not needed by this time, of the high test-retest reliability of the SVIB. The largest mean difference observed is but seven standard score points, and the variabilities of all distributions are quite small. Related to this relatively inconsequential test-retest variability is the obvious similarity of the in-chance and not-in-chance groups in the two samples, observed differences in most cases being in the range of one or two standard score points. These data, then, provide no basis upon which to question the many findings on high test-retest reliability of the instrument in question, on the grounds that SVIB profiles contain two different "kinds" of data.

Thus, while a given range on the raw score scale of any occupational scale may be interpreted as "nonsignificant," it does not seem reasonable to make this interpretation on the basis that the score arrived at that point through "chance" factors. The finding that the modal score for these subjects is a "chance" score is probably an artifact of the fact that most subjects do *not* have interests similar to the interests of a specific occupational criterion group.

The implications of these findings are at least threefold. Most important is the problem of how to interpret "gray areas" to, with, or about counseling clients. This is, basically,

the question: Why have gray areas on profile sheets? While there may be some utility in calling attention to the zero raw score point on each scale (which point approximates the chance mean), there would seem to be more value in omitting this shading altogether. While such an omission ignores the "paradox" that some men in an occupational criterion group have negative scores on their own scales (Strong, 1943, p. 85), the advantage gained is that all scores would be interpretable in the same manner, with the interpretation emphasizing the empirical derivation of the scale concerned. Also, there *is* error at all points on all scales. Thus, there seems to be no special advantage in calling attention to one special kind of error.

The second implication concerns the role of chance in pattern analysis of the SVIB. Pattern analysis by most methods, for example, Darley and Hagenah (1955) or Stephenson (1960), takes cognizance of the chance area as chance. If it develops that chance is not chance, then pattern analysis thinking must be revised to either incorporate the zero raw score concept or to utilize letter grade concepts only.

The final implication is that it is conceptually not at all satisfying, in considering vocational development theory, occupational choice theory, or vocational counseling technique, to think in terms of chance, which cannot be manipulated (by definition) and cannot be replicated (see the random numbers column of Table 1). This is especially disturbing in view of the amount of one's interest test data that must be discarded if one considers gray area chance as being in fact chance (a step that most investigators apparently do not take in practice).

Summary

Strong Vocational Interest Blank (SVIB) letter grade normative data are presented for three male groups with scores falling in the "chance" area considered as a separate letter grade category. While the proportions of scores in the chance area differ for the three groups, in all groups such proportions are the highest of any letter grade category. Using the test-retest data of King (1957) and Powers (1956), each subject's score on each of 44 SVIB occupational scales is tabulated as "in-chance" or "not-in-chance" on test. Retest deviations were computed in standard score units regardless of direction. These absolute average deviations were compared between King's two groups and between Powers' two groups. Comparisons were made over scales and over persons. All results were interpreted as implying that in-chance scores, being no more variable than not-in-chance scores, should probably not be considered as resulting from chance, as the term is normally used. Other implications of these findings were briefly discussed.

References

Darley, J. G., & Hagenah, Theda. *Vocational interest measurement.* Minneapolis: Univer. Minnesota Press, 1955.

King, L. A. Stability measures of Strong Vocational Interest Blank profiles. *J. appl. Psychol.*, 1957, **41**, 143–147.

Lyerly, S. B. "Chance" scores on the Strong Vocational Interest Blank for men. *J. appl. Psychol.*, 1957, **41**, 141–142.

Powers, Mabel K. Permanence of measured vocational interests of adult males. *J. appl. Psychol.*, 1956, **40**, 69–72.

Rand Corporation. *A million random digits with 100,000 normal deviates.* Glencoe, Ill.: Free Press, 1955.

Stephenson, R. R. Pattern analysis of the Strong Vocational Interest Blank for men. *Amer. Psychologist*, 1960, **15**, 455. (Abstract)

Stephenson, R. R. Predicting SVIB profiles of high ability male arts college freshmen. *Personnel guid J.*, 1961, **39**, 650–653.

Strong, E. K., Jr. *Vocational interests of men and women.* Palo Alto: Stanford Univer. Press, 1943.

Strong, E. K., Jr. *Manual for Strong Vocational Interest Blanks for men and women.* Palo Alto: Consulting Psychologists Press, 1959.

(Received December 23, 1960)

Journal of Applied Psychology
1961, Vol. 45, No. 6, 420–427

VISUAL TARGET LOCATION AS A FUNCTION OF NUMBER AND KIND OF COMPETING SIGNALS[1]

DAVID M. PROMISEL[2]
Johns Hopkins University

A common problem in the design of man-machine systems is that of presenting information to the man quickly and accurately. The problem may be to present one informational variable, i.e., information that varies along some single continuum, or several informational variables. One method of doing this consists of setting up a relationship between the values, or continua, of the informational variables and the variable aspects of a display. Visual displays are most useful for this purpose because of the number of ways in which a visual signal can vary. For example, the signals on a PPI are recognized as small lighted areas on the tube face. Informational variables can be presented to a man by assigning meaning to, or coding, the variable dimensions of these signals to represent the information.

There are several general methods of accomplishing this assignment. One method, which can be called the nonredundant method, is to relate an informational variable directly to some one variable dimension of the signal, e.g., the varying brightness of the above "small lighted area on the tube face" might be made to signify the size of the object the signal represents. Here brightness is the visual dimension of the signal, and size of the object is the informational variable. Theoretically, as many informational variables can be coded as there are variable dimensions of the signals. A large red lighted area might represent a large object at an altitude of 20,000 feet, a small blue lighted area might represent a small object at 15,000 feet, etc.

Another method is that of redundant coding, i.e., to signify an informational variable by a combination of signal dimensions each of which uniquely, as well as in combination, represents the informational variable. The value of the informational variable being presented can be determined by referring to any one of the several signal dimensions used as codes, e.g., a large red lighted area would now represent only the altitude of the object with both large and red signifying 20,000 feet.

A third technique is to use a partially redundant coding system. In this case the informational variable is again signified by a combination of dimensions but the value of this variable cannot be determined exactly from any one of the dimensions, as is true in the completely redundant case. At least one of the dimensions does uniquely determine the value but other dimensions merely limit the possible range of the values. For example, given a square blue signal representing the altitude of an airplane, the square could signify that the altitude was less than 20,000 feet while blue could indicate exactly 17,000 feet.

Dimensions that may be used as visual codes include: hue, geometric shape, area, number, length, angular orientation, brightness, flash rate, and stereoscopic depth. Each of these has its limitations in terms of the number of identifiable steps within the dimension, its effect upon operator fatigue and distractability, the interference with other dimensions, and the speed and accuracy of identification of the components, or values, of the dimension. There are also engineering problems involved in presenting any one of these codes.

There are many factors apart from the codes themselves which affect the efficiency of use of a visual coding system. One such factor may be termed "clutter"; the more complex the visual display the harder it is to find any specific signals (Eriksen, 1953,

[1] This research was done under Contract Nonr-248(55) between the Office of Naval Research and Johns Hopkins University. This is Report No. 15 under that contract. Reproduction in whole or in part is permitted for any purpose of the United States Government. The author wishes to express his appreciation to A. Chapanis and W. R. Garner for their guidance during this investigation.

[2] Now at Bell Telephone Laboratories, Whippany, New Jersey.

1954b; French, 1954; Green & Anderson, 1955, 1956; Reese, 1953). "Competition" is another factor which occurs when irrelevant signals on the display have some, but not all, dimensional values identical to the desired signal or target. For example, if large red signals were being searched for, a small red signal would be a "competing" signal. When performance in locating targets is deficient the use of redundant coding may be of value in improving it (Eriksen, 1952, 1953; Green & Anderson, 1955, 1956; Reese, 1953).

Most of the research done in this area has been concerned with the coding of one informational variable at a time and with the use of redundant coding. There has been little work dealing with the multiple coding of signals to present more than one kind of information simultaneously. This experiment was conducted to investigate the effect on operator performance with multicoded visual signals of the following variables: the number of targets searched for, the value of the targets searched for in terms of hue and shape, and the quantity and distribution of the competition on the display.

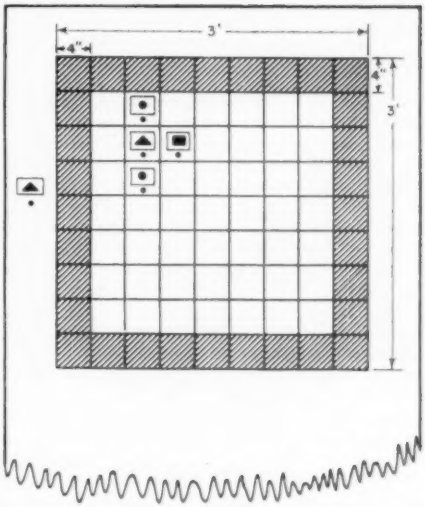


FIG. 1. Schematic of the display equipment. (The subject's task was to search through the 40 signals located in the inner—unshaded—49 squares of the display board and to touch a stylus to the discs under all those which matched the sample, hung at the left, in both hue and shape.)

METHOD

Materials and Apparatus

The two visual dimensions used as codes in this experiment were hue and shape. Individual values of these dimensions can be discriminated more quickly and accurately than can values of other dimensions (e.g., in general, a given hue can be picked out from other hues more quickly and with fewer mistakes than a given sized object can be picked out from objects of other sizes). As components of redundant codes, hue and shape, again, appear to be the best dimensions (Eriksen, 1952, 1953, 1954a; Reese, 1953).

The forms used as components of the shape dimension were circles, hexagons, diamonds, triangles, crosses, stars, and squares. These were selected from a list of shapes recommended for use in coding (Baker & Grether, 1954). The hues used were (in Munsell notation) R 5/6, YR 5/6, Y 5/6, GY 5/6, G 5/6, BG 5/6, and B 5/6. These represent the areas of most sensitive color vision on the jnd vs. wave length plot. The objects, or signals, consisting of all 49 combinations of the two codes, were cut out of Munsell paper and glued on to white, $3'' \times 2\frac{1}{4}''$ cards. To keep the size of the objects reasonably constant they were all made just circumscribable by a circle 1" in diameter.

A schematic of the display equipment is shown in Figure 1. The display board was 3' square, flat white, and ruled off by black lines into 81 squares, 4" on a side. The signal cards were hung on hooks located near the top of each of the inner 49 squares. Located below the card in each square was a $\frac{1}{2}''$-diameter circular metal disc. The display board was perpendicular to subject's (*S*'s) line of sight and was mounted in a black stand so that the lower edge of the board was 42" from the floor. A sample of the target signal was hung on the stand to the left of the display board. The stand also contained a sliding panel which, when raised, hid the display from *S*. Illumination was provided by four overhead cool white fluorescent bulbs.

Each of the discs on the display was connected to a different symbol on an IBM input-output typewriter. When a stylus, attached by a long connection to the right of the stand, was touched to a disc an electrical circuit was closed and the appropriate symbol was printed out by the typewriter. When the sliding panel dropped, it tripped a microswitch, starting an electric timer. The timer was stopped by *S* touching the stylus to a disc mounted on the stand below the sample target signal.

Subjects

Twenty male *Ss* were used. They were obtained from among the graduate and undergraduate students of Johns Hopkins University. All *Ss* passed a shortened version of the Ishihara Color Vision Test (Chapanis, 1948). The *Ss* were paid at an hourly

rate and rewards were also given for the best three performances.

Procedure

For each trial 40 signal cards and nine blank white cards were placed in random order on the display board. The values of these cards were selected from the 49 possibilities, subject to the constraints imposed by the experimental design. The sample target signal was put in place before the trial and remained there throughout the trial. The *Ss* were handed the stylus and were told that, when the sliding panel dropped, revealing the display and starting the timer, they were to touch the stylus to the discs under all the signals that matched the sample in both hue and shape. When they thought they had found all such signals they were to touch the stylus to the disc under the sample, thereby stopping the timer. *Ss* were given two sample trials but were given no instructions as to how to perform the task. The time score was the time between the dropping of the panel and the stopping of the timer. The error score was the number of omission and commission errors, i.e., the number of target signals not discovered and the number of nontarget signals mistaken for target signals. It was explained to the *Ss* that performance would be measured by ranking the *Ss*' total time scores and error scores separately and then summing these two ranks for each *S*, giving them equal weight. The lowest sum of the ranks would indicate the best performance by an *S*, for the purpose of awarding the prizes.

Experimental Variables

The experiment investigated the effects on performance of four independent variables.

Variable VT was the value of the target signal. Five of the 49 hue-shape combinations were used as targets. These were the R circle, the YR hexagon, the GY triangle, the BG star, and the B square.

Variable NT was the number of target signals to be located on the display. These numbers were unknown to the *Ss* and were 0, 2, 4, 6, or 8. Inasmuch as the total number of signals on the board was held constant at 40 an increase in the number of target signals always corresponded to a decrease in the number of nontarget signals.

Variable NC was the number (0, 8, 16, 24, or 32) of competing signals on the display. These were signals identical to the target signal in either hue or shape but not both. Variable NC was, of course, independent of Variable NT but an increase in competing, nontarget signals was of necessity confounded with a decrease in noncompeting, nontarget signals (i.e., those signals with neither the same shape nor hue as the target).

Variable DC represented the manner of distribution of the competition between the hue and shape dimensions. Under condition DC_1 half the competing signals competed in terms of hue and half in terms of shape. Under condition DC_2 all the competition was in the hue dimension.

Experimental Design

The 20 *Ss* were randomly assigned to two blocks of 10 *Ss* each. These two blocks were treated identically. Within a block the *Ss* were again divided randomly into two groups of five each. One group was given condition DC_1, the other condition DC_2, but otherwise they, too, were treated identically.

Within each DC condition in each block the five *Ss* received combinations of the remaining three variables, VT, NT, and NC. Each of these variables had five levels. Every *S* was given 25 trials, representing a different latin square combination of these three variables. These latin squares were formed in such a way that a design containing the *Ss* and any two of the three variables also formed a latin square for each level of the third variable. This procedure produces a complete factorial arrangement within the DC conditions except that some of the interaction terms are confounded with subject variances. Comparable *Ss* in the two blocks, although receiving the same variable combinations, received these in a different random order.

Results

The two dependent variables in this experiment were time scores and error scores. The time scores were subjected to an analysis of variance for a modified random block design with fixed effects. The distribution of these scores showed some mild skewness and heterogeneity but to transform the data promised both to be of little value and to complicate the interpretation of the data. Thus no transformation was made. The results of the analysis showed the VT, NT, NC, and NC × DC terms to be significant.

There were not enough errors, 129 in 500 trials, to justify the use of parametric statistical tests of them. The 25 commission errors were not subjected to any statistical analysis at all. The remainder of the errors, 104, were omission errors, i.e., there were 104 target signals that were not discovered. Using total omission errors for each *S* as the unit of measure, the difference between DC_1 and DC_2 was not significant according to the Mann-Whitney *U* test. Based on two-way tables of the omission error scores for the Variables NT, NC, and VT, and assuming no interactions, only the NT and VT variables are significant using the Friedman two-way analysis of variance.

Variable VT

Figures 2 and 3 show the relative effect on performance of the five different hue-shape

combinations used as targets. It should be noted that, apart from the reversal between YR hexagons and B squares and the steeper slope of the omission error curve, target value affects performance in a similar manner whether it is measured in terms of task time or in terms of the number of omission errors. Figure 2 shows that, when all other conditions are averaged out, the R circle results in the quickest task time and the GY triangle the slowest.

These results with regard to target value are consistent, but in themselves they give no indication of the critical factor which made one target value better than another. Since with the targets actually used hue and shape were completely confounded, it is impossible to know whether the R circle was best because it was a circle or because it was red. It is possible, however, to derive some reasonable inferences about which of these two factors is critical by examining some of the subconditions—subconditions which vary in the extent to which each of these two codes was required to be used by *S*.

Table 1 gives the rank-orders of the time scores for the different target values for four subconditions. Under the NC_0 condition there are no competing signals on the display. *S* may identify the targets in terms of their hue or shape or both. The rank-order of the times under this subcondition is identical to the

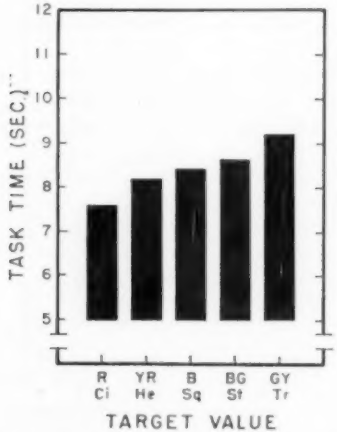


FIG. 2. Time to locate different hue-shape combinations, averaged over all experimental conditions.

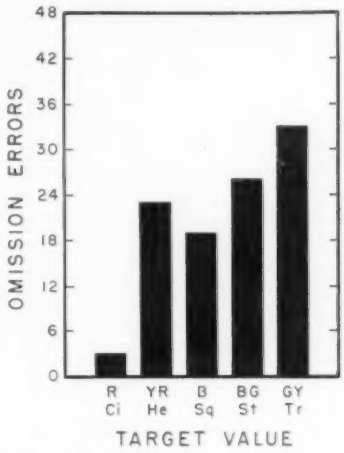


FIG. 3. Total omission errors made while searching for different hue-shape combinations.

overall rank order as shown in Figure 2. Under Subcondition DC_1, target location is dependent on both hue and shape, and in this case, the rank-order of the times is still the same. With the DC_2 condition, identification must be made in terms of shape, but hue is still helpful to some extent, and with this condition the rank-order of times is still correlated with those of Figure 2, but not perfectly. Finally, under one condition ($NT_8NC_{32}DC_2$) there are eight targets and the remaining 32 signals all compete in the hue dimension. Thus with this subcondition, the identification must be made on the basis of shape alone, and in this case the rank-order of times is completely uncorrelated with those in Figure 2. It seems reasonable, therefore, to suppose that the rank-order of the times shown

TABLE 1

RANK-ORDER OF TIME SCORES FOR THE DIFFERENT TARGET VALUES

Target values	R Ci	YR He	B Sq	BG St	GY Tr
Shape and/or hue dependent	1	2	3	4	5
Shape and hue dependent	1	2	3	4	5
Shape dependent, hue helpful	1	4	3	2	5
Shape dependent, hue not helpful	4	5	1	2	3

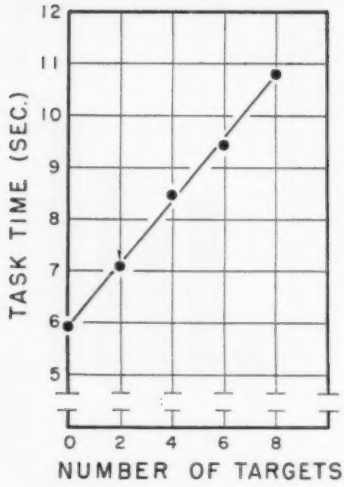


FIG. 4. Average time to locate all the target signals as a function of the number of target signals. (Each point is the average for all other experimental conditions.)

in Figure 2 is caused primarily by differences in hue and not by differences in shape.

The only difference between the subcondition in which target identification is dependent completely on shape and the other three subconditions is the increasing dependency of the identification on the hue dimension. There is a shift in the rank-order away from that of the purely shape dependent subcondition as the dependency on hue increases. It is possible to verify further this description of the shift. Of the 25 commission errors, 23 occurred under the DC_1 condition. Of these, 15 occurred when G stars were mistaken for BG stars and six occurred when G triangles were mistaken for YG triangles. In other words, the confusion of both the BG and YG hues with the G hue caused almost all of the commission errors. Also, *S*s commented to the effect that hexagons and circles were relatively difficult to pick out, as were YG and BG. These facts concerning relative performance with the different code values are mirrored in the rank-orders mentioned above. Thus, the characteristics of the hue code used in this experiment override those of the shape code in determining task time.

Whether conclusions regarding the codes used here can be extended to hue-shape codes in general is an open question. The lack of significance of any of the interactions involving the VT variable indicates, however, that the results pertaining to the other independent variables that were obtained with these specific codes can be stated with some generality.

Variable NT

Figures 4 and 5 show task time and omission errors, respectively, as a function of the number of target signals on the display. The task time function is described as a straight line, justified by an analysis of the data into orthogonal polynomial terms that shows only the linear component to be significant. Although no such analysis was possible for the omission error function, it too appears to be linear increasing over at least the first four experiemental levels. That the drop in the function when there are eight target signals on the display is meaningful is suggested by the notion that if the value of NT were carried to the extreme where all of the display signals were targets, then there would probably be no omission errors. This drop might indicate the beginning of a decreasing function.

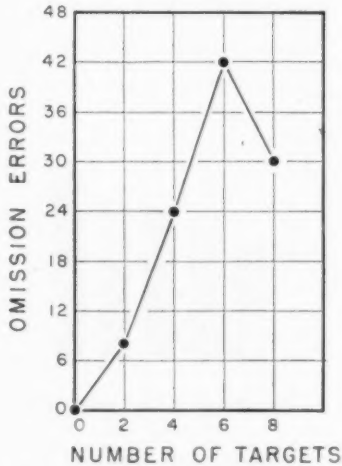


FIG. 5. Total omission errors made as a function of the number of targets to be found.

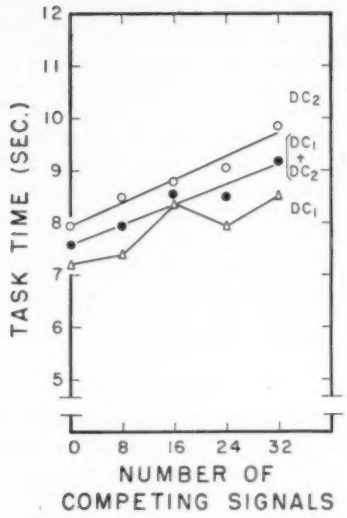


FIG. 6. Average time to locate targets as a function of the number of competing signals. (Each point is the average for all other experimental conditions. The function for the DC_1 condition has been drawn connecting the points because it contains a quartic component as well as a linear component.)

Variables NC and DC

The effect of the number of competing signals on task time is shown in Figure 6 for the overall situation and that occurring when DC_1 and DC_2 are analyzed separately. The linear functions for the overall condition and the DC_2 condition are again justified by the significance of the linear terms whereas the analysis shows a significant quartic as well as linear component for the DC_1 condition. This quartic component indicates that there is some essential difference in the task required by the latter condition, a multicode situation, as compared with the former condition, a partially redundant code situation. This task difference does not result in a significant difference in task time between the two conditions, however. This lack of significance is caused by the fact that the proper error term for testing the significance of this difference is a between-subjects term, which was quite large in this experiment. It seems likely that with a larger number of *Ss* a significant difference would be found.

Overall performance

Ss' performance was measured in terms of both task time and errors. The data show that where the independent variables affected the two measures, they did so in a similar manner for both of them. This was despite a tendency for the *Ss* to trade off speed for accuracy and vice versa. The Spearman rank correlation coefficient for the relationship between the total time and error scores for each of the 20 *Ss* is -0.626. The similarity between the two measures is apparent even though the *Ss*' trade-off tendencies serve to mask it out.

DISCUSSION

Ss' task in this experiment required them to make two kinds of responses, overt and covert. The overt response is that actually required in tapping the metal discs, and the covert response is the discrimination required to decide which discs to tap. The NT variable was the only one affecting the overt responses, and, in terms of effect on performance time, it was by far the most important. The sum of squares in the analysis of variance for this variable was far greater, by a factor of 10, than those for the main effects of the other variables and amounted to almost a third of the total sums of squares. It seems likely that the linear effect of NT on task time is due primarily to the linear increase in the number of overt responses required as the number of target signals increases. The nonzero intercept of the function is created by the necessity for *S* to scan the display even when there are no targets on it. It follows that the most substantial savings in task time in this experiment could have been effected by designing *S*'s task of designating targets to require fewer and/or quicker overt responses.

Variables VT, NC, and DC involved only covert, or discriminative, responses from *S* and had relatively small but nevertheless significant effects on task time. Thus, the NT effect is probably a composite of two functions. One is the effect produced on task time by NT as a result of the overt responses required by it; the other is the result of the covert responses required. Since the overt responses would be far more time consuming

than the covert responses, the former would mask out the latter. Perhaps the covert response function for NT is similar in shape to the omission error vs. NT curve, since the number of omission errors is affected only by covert responses and not at all by overt responses.

The NC variable affected only the difficulty of the perceptual discriminations, not the number of them, i.e., the amount of clutter on the display remained constant. For any value of NT *S* had to examine a constant number of nontarget signals. But as the value of NC increased, the difficulty of discriminating the target signals from the nontarget signals increased. If the significance of the NC × DC term is to be given meaning it must be that the difficulty of discrimination is a function not only of the quantity of competing targets, but also of their quality.

A further explanation of this function can be given in terms of the "effective" number of signals to be examined as determined by the manner of distribution of the competition. In the partially redundant case, DC_2, it has been shown (Green & Anderson, 1955, 1956) that the competing value effectively determines the area of search. *S*, in some sense, only examines those signals having the competing, i.e., partially redundant, value. This fact would explain the linear effect obtained in DC_2: the larger this effective area, as NC increases, the longer the search time. Now, if one assumes that for sufficiently large numbers of competing signals one of the two codes in DC_1 approximates a partially redundant code and thus determines the effective search area, whereas for smaller numbers the number of competing hue and shape signals combined determines that area, it is possible to generate a predicted curve with a shape similar to the actual curve shown in Figure 6 for DC_1. The predicted curve is based on the *S*s' having to search out effective areas of the following numbers of signals (averaged over NT) as NC increases from 0 to 32: 4, 12, 20, 16, 20. Thus, at the lower three levels of NC the effective area to be searched by the *S* includes the targets (an average number of four) and all the competing signals, in both hue and shape. For example, at the second level of NC there is an average of four targets on the display, four signals competing in hue and four competing in shape, producing an effective search area of 12 signals. At the upper two levels, the effective area of search would include, again, the targets plus only those signals which compete in one dimension (perhaps the signals that compete in hue). At the highest level, for instance, there are the four targets, 16 competing hue signals and 16 competing shape signals. It is hypothesized, though, that *S*, in effect, picks one of the target values (e.g., target hue) and then limits his area of search to only those signals having that value.

Summary

Performance in obtaining information from a double coded visual display was investigated as a function of these variables: the number of target signals on the display, the values of the hue-shape combinations used as target signals, the number of competing signals on the display, and the distribution of this competition between the hue and shape dimensions.

The number of targets, the value of the targets, and the number of competing signals all significantly affect task time—the number of targets having by far the greatest effect. Both task time and errors are greater for some targets than for others, and analysis suggests that hue rather than shape is the critical factor. Both task time and errors are again larger for larger numbers of targets, and task time alone increases with the larger number of competing signals. While there is no difference in task time between locating double coded targets and partially redundant targets, the tasks themselves appear to be different.

REFERENCES

Baker, C., & Grether, W. Visual presentation of information. *USAF WADC tech. Rep.*, 1954, No. 54-160.

Chapanis, A. A comparative study of five tests of color vision. *J. Opt. Soc. Amer.*, 1948, **38**, 626-649.

Conover, D., & Kraft, C. The use of color in coding displays. *USAF WADC tech. Rep.*, 1958, No. 55-471.

Duncan, A. *Quality control and industrial statistics.* Chicago: Richard D. Irwin, 1959.

Eriksen, C. W. Location of objects in a visual display as a function of the number of dimensions on

which the objects differ. *J. exp. Psychol.*, 1952, **44**, 56–60.

ERIKSEN, C. W. Object location in a complex perceptual field. *J. exp. Psychol.*, 1953, **45**, 126–132.

ERIKSEN, C. W. Multidimensional stimulus differences and accuracy of discrimination. *USAF WADC tech. Rep.*, 1954, No. 54-165. (a)

ERIKSEN, C. W. Partitioning and saturation of the perceptual field and efficiency of visual search. *USAF WADC tech. Rep.*, 1954, No. 54-161. (b)

FISHER, R. A., & YATES, F. *Statistical tables for biological, agricultural and medical research*. London: Oliver & Boyd, 1953.

FRENCH, R. S. Pattern recognition in the presence of visual noise. *J. exp. Psychol.*, 1954, **47**, 27–36.

GREEN, B., & ANDERSON, L. Size coding in a visual search task. *Mass. Inst. Technol. group Rep.*, 1955, No. 38-17.

GREEN, B., & ANDERSON, L. Color coding in a visual search task. *J. exp. Psychol.*, 1956, **51**, 19–24.

HUMPHEY, G. *Thinking: An introduction to its experimental psychology*. New York: Wiley, 1951.

REESE, E. P., REESE, T. W., VOLKMANN, J., & CORBIN, H. H. Psychophysical research summary report: 1946–1952. *USN Spec. Dev. Cent. tech. Rep.*, 1953, No. 131-1-5.

SCHEFFE, H. *The analysis of variance*. New York: Wiley, 1959.

(Received January 12, 1961)

Journal of Applied Psychology
1961, Vol. 45, No. 6, 428-430

EFFECT OF PERSONAL CHARACTERISTICS ON RELATIONSHIPS BETWEEN ATTITUDES AND JOB PERFORMANCE[1]

FRANCIS D. HARDING AND ROBERT A. BOTTENBERG

Personnel Laboratory, Lackland Air Force Base

Because the relationship between measurements of attitudes and job performance has generally been found to be low, the nature of research in this area has been undergoing change. One conclusion has been to no longer consider satisfaction and morale indexes as variables intervening between supervisory and organizational characteristics on the one hand and productivity on the other (Kahn, 1960). Another suggestion has been that morale be considered as an output variable, and attempts be made to delineate the conditions under which morale is related to other group outcomes (Whitlock & Cureton, 1960).

This study investigates the possibility that by giving more consideration to characteristics of workers and jobs, greater relationships between attitudes and productivity will be found. In this regard, it has been shown that workers performing different kinds of work tend to react differently toward management policies (Sayles, 1958). Also, that the amount of work experience of an individual and the organization to which he belongs does affect his ratings of the work activities making up his job (Harding & Naurath, 1960).

This report deals with a reanalysis of the data collected by Whitlock and Cureton in their investigation aimed at developing effective means for measuring morale among Air Force personnel. In that study no attempt was made to control for the influence upon an airman's attitude of his rank, age, length of service, career intention, or type of work performed. The purpose of the present research is to determine the extent to which such biographical data affect the relationships between attitude and job performance. It was hypothesized that personal characteristics such as rank, length of service, career intention, type of work, and months within the organization when combined with airmen's scores on attitude scales would contribute to the prediction of criteria such as job performance and absences from work.

[1] The research reported in this paper was sponsored by Personnel Laboratory, Aeronautical Systems Division, under AFSC Project 7719, Task 17130.

PROCEDURE

The study involved the application of multiple regression techniques to the data collected by Whitlock and Cureton. Attitude scores and biographical data variables were used to predict criteria of job performance for 376 airmen. First, all predictors were allowed to enter the regression equation. Then, by removing certain groups of predictor variables it was possible to determine their independent contribution over and above that of the variables remaining in the equation.

The variables used in the regression analysis may be classified as follows:

Attitude variables:

1. Satisfaction with supervisor—supervisor-subordinate relations
2. Satisfaction with Air Force as a military organization—how well the Air Force is felt to be accomplishing its mission
3. Job satisfaction—interest and pride in job
4. Satisfaction with civilian community—reaction to relations between local community and airmen
5. Satisfaction with Air Force life—overall benefit of being in the Air Force
6. Satisfaction with management and communication—how well informed the airmen feels he is
7. Satisfaction with unit—pride in work group and opinion of the group's leaders
8. General morale—covering everything that is measured by the other scales

Biographical variables:

1. Rank of the airmen—this characteristic was represented in the regression analysis by five categorical variables, one each to indicate membership in a particular rank category
2. Career intention—whether or not an airman intended to make a career of Air Force service
3. Length of service—the sample of airmen was grouped into those who had less than four years of service and those who had more service
4. Kind of work performed—the airmen were divided into two groups: those whose work was closely

TABLE 1

MULTIPLE R^2s OBTAINED IN PREDICTING THE JOB PERFORMANCE AND ABSENCES CRITERIA

Predictors	Criterion		
	Job performance rating	Job performance ranking	Estimated absences
All variables	.18	.16	.14
Biographical variables only	.08	.09	.04
Attitude-rank interaction variables removed	.11	.14	.07
Linear attitude variables only	.07	.07	.04

related to flying and maintaining aircraft and those whose work was less involved in the actual operational mission

5. Number of months in squadron—the number of months the airmen had been assigned to his present unit

Interaction variables:

1. Interaction between rank and attitude—variables which show the scores for a particular attitude of airmen who hold the same rank. For each of the five airmen ranks there are eight variables, one for each of the attitude scales. The elements of such interaction variables are the attitude scores of airmen holding the same rank and zeros for all airmen holding other ranks

Criterion variables:

1. Job performance rating—each airman was rated on a six-point scale by his supervisor
2. Job performance ranking—each airman was ranked in respect to overall job performance by his supervisor
3. Estimated absences—the number of times each airman had asked to be excused from duty during previous 3-month period for reasons other than his own sickness

RESULTS AND DISCUSSION

How much of the criterion variance is accounted for by various combinations of predictors is shown in Table 1 by use of multiple R^2s. The first row shows the proportion of variance accounted for when all the predictors are allowed to enter the equation. The values represent correlations of .43, .41, and .38, respectively, for the three criteria. The size of the correlations are encouraging and suggest that with additional refinements the variables would provide usable research tools.

An analysis was performed to test whether biographical variables would contribute a statistically significant amount of unique variance when used with a linear combination of the attitude variables. This was shown to be the case for the two job performance criteria. These findings are in keeping with the hypotheses stated above and mean that if you are only using attitude variables to predict job performance, you can improve your prediction by adding biographical variables such as those used in this study.

An additional analysis was made to determine if the attitude variables added to the prediction of job performance beyond that which was predicted by only the biographical variables. For example, in the present study, it is possible that an airman's rank, length of service, career intention, or the type of work performed might account for the variance in job performance to the extent that attitude variables added nothing further. As shown in Table 1, the multiple R^2, obtained when only biographical data variables are used as predictors is .08 for job performance ratings and .09 for the job performance rankings, and for the criterion of estimated absences it is .04. When variables which reflect attitude measures are added to the group of predictors the squared multiples become .18, .16, and .14 for the three criteria. While appreciable, these increases in the multiple R^2s are not statistically significant at the .10 level. One possible interpretation is that the increase in predictive efficiency is due to the use of a very sizeable number (40) of additional predictor variables rather than the addition of independent variance. The failure of the attitude variables to make a significant independent contribution beyond that which is contributed by the biographical variables is an

important finding. It suggests that much of the variance which in the past has been attributed to attitudes is held in common by the more easily obtained biographical variables.

The extent to which the attitude by rank interaction variables contribute to the prediction of job performance was also tested by a comparison of the appropriate multiple R^2s. For each of the three criteria there was an increase in the proportion of predictable variance when the interaction variables were added to biographical and linear attitude predictors, however, none were large enough to be statistically significant.

Conclusions

The suggestion was made that one reason for the lack of relationship usually found between attitudes and job performance is the failure to take into account important biographical and situational variables. It was hypothesized that the inclusion of such variables in multiple regression problems would improve the prediction of job performance.

Comparisons of the multiple R^2s, obtained when several combinations of variables were used to predict criteria of job performance and absences from job permit the following statements to be made:

1. The contribution of attitude variables to the prediction made by biographical variables is not significant.

2. The contribution of attitude–rank interaction variables to the combination of linear attitude variables plus biographical variables is not significant.

3. The contribution of biographical variables to a linear combination of attitude variables is significant for job performance criteria but not for the estimated absences criterion.

These findings suggest that much of the criterion variance associated with attitude measures is also related to more easily obtained biographical characteristics of workers and that biographical variables should be considered when relating attitudes to productivity.

References

Harding, F. D., & Naurath, D. A. Effects of job experience and organization on the rating of tasks. *Engng. industr. Psychol.*, 1960, **2**, 63–68.

Kahn, R. L. Productivity and job satisfaction. *Personnel Psychol.*, 1960, **13**, 275–287.

Sayles, L. R. *Behavior of industrial work groups: Prediction and control.* New York: Wiley, 1958.

Whitlock, G. H., & Cureton, E. E. Validation of morale and attitude scales. *USAF WADD tech. Rep.*, 1960, No. TR-60-76.

(Received January 13, 1961)

Journal of Applied Psychology
1961, Vol. 45, No. 6, 431–434

THE STABILITY OF GUILFORD–ZIMMERMAN PERSONALITY MEASURES

JAY M. JACKSON

University of Kansas

In using test scores in the selection of employees, it is desirable that so-called measures of personality remain relatively stable over time. This is the only justification for assuming that they describe attributes of the *person*, rather than the person's reaction to his immediate *situation*. Many persons, for example, will feel somewhat insecure in a new environment with strangers. But some persons will have a greater tendency than others, and will require more supportive and less threatening situations. One would expect such persons to be less secure than the average regardless of situational changes, and their relative insecurity to persist over time.

As part of a larger study[1] of factors related to turnover in two commercial offices of a telephone company, the Guilford-Zimmerman Temperament Survey, a standardized personality inventory originating in factor analysis, was administered to 96 female employees. A number of findings pointed to the relevance for performance of several of the traits measured by the GZTS. It was found, for example, that girls who were relatively high on Sociability and Ascendance had a greater aptitude for the job, and those low on Security, an index constructed by the investigator from items of the inventory (Jackson, 1956), had considerable difficulty making an adjustment to the work situation. In the light of these and other results it was concluded that the GZTS showed promise as a potential management aid in selection of personnel for the job being studied, *if it could be concluded that the test was measuring relatively persistent characteristics of persons*, rather than reflecting reactions to the test situation, the current life situation, or temporary feeling states.

[1] A number of papers reporting results of this project are in the process of publication. A report was made to the company in 1955 (see Jackson, 1955).

An examination of the quite extensive literature yields no evidence regarding the question raised: the GZTS has been correlated with intelligence scores (Bass, Wurster, Doll, & Clair, 1953; *Psychometric Notes*, 1954), college grades (*Psychometric Notes*, 1954), academic achievement (Bendig & Sprague, 1954; Goedinghaus, 1954), leaderless group discussion scores (Bass, McGehee, Hawkins, Young, & Gebel, 1953; Bass et al., 1953), success in training occupational therapy students (Booth, 1957), personnel assessments (Hilton, Bolin, Parker, Taylor, & Walker, 1954), selection and placement of engineers (Kirkpatrick, 1956), perseveration of set (Pitcher & Stacey, 1954), and other personality tests (Gilbert, 1950; Jones, 1954; Kaess & Witryol, 1957). Norms have been reported for males vs. females (Eisele & Cottle, 1951; Isaacson & Cottle, 1952), urban vs. rural (Isaacson & Cottle, 1952; Johnson & Cottle, 1952), Negroes vs. whites (Johnson & Cottle, 1952), and normals vs. neurotics (Franks, 1956). But there appears to be no published evidence regarding the relative stability of the trait scores over time.

Method

The Guilford-Zimmerman Temperament Survey was administered in November 1953 to 72 female white-collar employees, called "Service Representatives," and 24 female supervisors in two different districts of a telephone company. A second administration took place in June 1955 after a lapse of approximately 18 months. By that time the original test sample had been reduced by turnover to 34 Service Representatives and 15 supervisors. In a year and a half, the life situations of many of the subjects in the original sample had changed considerably. By the time of the second administration, all subjects were considerably more experienced, on-the-job training had been completed, some subjects had been married, some promoted, and a few demoted. For many of the subjects, friends in the office had left and new friendships had been formed; their supervisors had changed at least once; and factors in their personal, emotional, and social lives could be

assumed to have undergone change. Thus it was assumed that the psychological situations for most subjects were different for the two administrations of the test.

The rationale provided the subjects for administering the personality inventory was that the research was attempting to determine whether a particular type of girl was attracted to their job, or whether they were simply a "cross section of American girls." It was implied that they were successful employees, and that management wanted to be able to select others like them. It appeared that these instructions were successful in reducing the threat for subjects. They were cooperative, relaxed, and the test situation seemed close to ideal.

Results

Scores for the two administrations of the GZTS are reported in Table 1. Since it is possible that length of employment at the time of initial testing might be a factor in whether or not changes occurred, results are presented separately for less experienced and more experienced employees. Means and standard deviations are also presented for the total sample, including supervisors. There is a striking similarity between the 1953 and 1955 mean scores for all scales and for each category of the sample. The difference between

TABLE 1

Stability of Guilford–Zimmerman Scores over an 18-Month Period for 49 Female Office Workers

Scale		Less experienced[a]		More experienced[b]		Total sample[c]	
		M	*SD*	*M*	*SD*	*M*	*SD*
General Activity	1953	16.9	4.6	19.1	6.9	18.2	6.2
	1955	15.4	5.5	20.0	6.0	17.9	6.2
Restraint	1953	14.6	5.1	16.8	4.1	16.9	4.8
	1955	14.3	5.3	17.0	4.4	17.0	5.1
Ascendance	1953	13.0	5.1	13.9	4.9	14.2	5.0
	1955	14.2	5.2	14.5	6.1	14.7	5.5
Sociability	1953	22.5	4.6	21.0	5.4	22.0	5.5
	1955	22.0	4.3	21.0	6.4	21.6	5.8
Emotional Stability	1953	18.0	4.4	20.2	4.7	20.0	4.7
	1955	18.6	4.2	20.9	5.4	20.4	5.4
Objectivity	1953	19.4	2.6	19.7	3.6	19.6	4.5
	1955	19.3	4.1	19.5	4.5	19.6	4.7
Friendliness	1953	16.4	3.6	17.0	6.2	17.7	5.4
	1955	15.6	3.8	15.8	5.0	16.7	4.9
Thoughtfulness	1953	16.9	6.0	17.4	4.1	18.7	6.6
	1955	16.5	7.2	16.4	4.5	17.3	5.6
Personal Relations	1953	20.3	4.4	21.0	5.5	21.6	5.4
	1955	20.8	4.8	20.7	5.6	21.6	5.8
Masculinity-Femininity	1953	10.9	3.3	11.6	4.4	11.4	3.8
	1955	12.1	3.6	10.4	4.9	11.7	4.3
Security[d]	1953	7.4	1.1	7.4	2.0	7.3	1.6
	1955	7.4	1.4	7.4	1.6	7.4	1.7

[a] Subjects with less than one year's service at initial testing, $N = 11$.
[b] Subjects with one year or more service at initial testing, $N = 23$.
[c] Includes supervisors, $N = 49$.
[d] This is a 10-item scale (see Jackson, 1956).

TABLE 2

PEARSON PRODUCT–MOMENT CORRELATIONS BETWEEN TWO ADMINISTRATIONS OF THE GUILFORD-ZIMMERMAN TEMPERAMENT SURVEY, 18 MONTHS APART, FOR 49 FEMALE OFFICE WORKERS

Scale	Less experienced[a]	More experienced[b]	Total sample[c]
General Activity	.75	.87	.77
Restraint	.70	.63	.75
Ascendance	.83	.71	.71
Sociability	.91	.91	.92
Emotional Stability	.31	.75	.69
Objectivity	.71	.63	.71
Friendliness	.59	.65	.68
Thoughtfulness	.92	.81	.82
Personal Relations	.37	.78	.68
Masculinity-Femininity	−.25	.75	.52
Security[d]	.21	.57	.50

[a] With $N = 11$, r must be .60 for the .05 level of significance, and .74 for $p = .01$.
[b] With $N = 23$, r must be .53 for the .01 level of significance.
[c] With $N = 49$, r must be .36 for the .01 level.
[d] This is a 10-item scale. Using the Spearman-Brown prophecy formula, the correlation for the total sample for a 30-item Security scale would be .75.

each pair of means was evaluated, using the customary t test for matched groups. In no case was it large enough to warrant confidence that any real change had taken place. Significance levels ranged between .70 and 1.00, far from approaching an acceptable level of confidence.

Even though mean scores on the test did not change significantly in the 18-month period, it is still possible that some scores increased and others decreased, with a compensatory effect that left the means unchanged. To explore this possibility, Pearson product-moment correlations were computed between the 1953 and 1955 scores. The results of this analysis are presented in Table 2.

Correlations for the total sample and for the more experienced subjects are all so large that they would occur by chance much less often than once in a hundred times. On a few of the scales for the less experienced group, correlations are relatively small, indicating considerable shifting in both directions from the original scores. Considering that there are only 11 persons in this category, this result does not appear to be too serious. Further examination of these scores indicated that the small correlations resulted from large changes in the scores of two subjects who had been married in the interval between the two test administrations. The correlations demonstrate substantial stability of the Guilford-Zimmerman scores, and in most cases are of the same order of magnitude as the reliability figures published in the GZTS manual (Guilford & Zimmerman, 1949).

CONCLUSIONS

The findings from this study indicate that within the limitations of the type and number of subjects involved, and to the degree that the psychological situations were different at the times of the two administrations, the Guilford-Zimmerman scale scores measure relatively persistent attributes of the persons tested. In this sense the test may be considered a *personality* inventory. It may be that when extreme changes occur in the life situation of a subject, such as marriage for a quite young girl, some personality traits such as Emotional Stability, Femininity, and Security, may be markedly affected. On the whole, however, the scores demonstrate considerable stability over time, and high test-retest reliability.

REFERENCES

BASS, B. M., MCGEHEE, C. R., HAWKINS, W. C., YOUNG, P. C., & GEBEL, A. S. Personality variables related to leaderless group discussion behavior. *J. abnorm. soc. Psychol.*, 1953, **48**, 120–128.

BASS, B. M., WURSTER, C. R., DOLL, P. A., & CLAIR, D. J. Situational and personality factors in leadership among sorority women. *Psychol. Monogr.*, 1953, **67**(16, Whole No. 366).

BENDIG, A. W., & SPRAGUE, J. L. The Guilford-Zimmerman Temperament Survey as a predictor of achievement level and achievement fluctuation in introductory psychology. *J. appl. Psychol.*, 1954, **38**, 409–413.

BOOTH, M. D. A study of relationship between certain personality factors and success in clinical training of occupational therapy students. *Amer. J. occup. Ther.*, 1957, **11**, 93.

EISELE, M. C., & COTTLE, W. C. The Guilford-Zimmerman Temperament Survey: I. With rural high school students. *U. Kans. Bull. Educ.*, 1951, **6**, 12–15.

FRANKS, C. M. Conditioning and personality: A study of normal and neurotic subjects. *J. abnorm. soc. Psychol.*, 1956, **52**, 143–150.

GILBERT, C. The Guilford-Zimmerman Temperament Survey and certain related personality traits. *J. appl. Psychol.*, 1950, **34**, 394–396.

GOEDINGHAUS, C. H. A study of the relationship between temperament and academic achievement. Unpublished master's thesis, University of Southern California, 1954.

GUILFORD, J. P., & ZIMMERMAN, W. S. *The Guilford-Zimmerman Temperament Survey: Manual of instructions and interpretations*. Beverly Hills, Calif.: Sheridan Supply, 1949.

HILTON, A. C., BOLIN, S. F., PARKER, J. W., JR., TAYLOR, E. K., & WALKER, W. B. The validity of personnel assessments by professional psychologists. *J. appl. Psychol.*, 1954, **38**, 287–293.

ISAACSON, L. E., & COTTLE, W. E. The Guilford-Zimmerman Temperament Survey: II. Urban high school students. *U. Kans. Bull. Educ.*, 1952, **6**, 46–50.

JACKSON, J. M. Summary of conclusions from the commercial department study. Ann Arbor, Mich.: Research Center for Group Dynamics, 1955. (Mimeo)

JACKSON, J. M. A Guilford-Zimmerman scale for psychological insecurity. Lawrence, Kan.: Author, 1956. (Ditto)

JOHNSON, A., & COTTLE, W. E. The Guilford-Zimmerman Temperament Survey: III. With urban Negro high school students. *U. Kans. Bull. Educ.*, 1952, **6**, 75–80.

JONES, M. B. Aspects of the autonomous personality: IV. Traits from the Guilford-Zimmerman Temperament Survey. *USN Sch. Aviat. Med. res. Rep.*, 1954, No. NM 001058.25.16.

KAESS, W. A., & WITRYOL, S. L. Positive and negative faking on a forced-choice authoritarian scale. *J. appl. Psychol.*, 1957, **41**, 333–339.

KIRKPATRICK, J. J. Validation of a test battery for the selection and placement of engineers. *Personnel Psychol.*, 1956, **9**, 211–227.

PITCHER, B., & STACEY, C. L. Is Einstellung rigidity a general trait? *J. abnorm. soc. Psychol.*, 1954, **49**, 3–6.

Psychometric notes, Number 5. Beverly Hills, Calif.: Sheridan Supply, 1954.

(Received January 27, 1961)

Journal of Applied Psychology
1961, Vol. 45, No. 6, 435–440

RELATIONSHIPS BETWEEN INDIVIDUAL PROFICIENCY AND TEAM PERFORMANCE AND EFFICIENCY

WILLIAM M. WIEST, LYMAN W. PORTER, AND EDWIN E. GHISELLI

University of California, Berkeley

Several studies have been reported comparing individual proficiency in the performance of a task with team proficiency. In general these studies have been addressed to one or another of the following three problems: (*a*) How does the level of performance of a team compare with the performance expected on the basis of a simple additive combination of proficiency of the individuals composing the team? (*b*) How well can the performance of a group of individuals working cooperatively on a task be predicted from knowledge of the proficiency of individual members of the team? (*c*) Is team performance better predicted from the individual performance of the more proficient members of a team or from the less proficient members, and how can the measures of individual proficiency best be combined to predict team performance?

Lorge, Fox, Davitz, and Brenner (1958) cite several studies indicating that the summation of individual contributions in the solving of cognitive problems exceeds the performance which these same individuals are able to achieve when working as a group. Thus in some cases "group interaction may inhibit the fullest potential contribution by its members" (p. 354). Studies by Husband (1940), Marquart (1955), Shaw (1932), Taylor and Faust (1952), and Watson (1928) support this generalization.

The second question has been pursued by Comrey (1953), Comrey and Deskin (1954a, 1954b), and Comrey and Staats (1955) in a series of experiments using both motor and cognitive tasks. With the motor task Comrey found that a surprisingly small proportion of the variance of the team performance was predictable on the basis of the proficiency achieved by the members when performing the same task individually. In all of these experiments using a motor task the proportion of predicted team variance was less than one half. But with a cognitive task Comrey found that considerably more than half of the team variance was predictable on the basis of individual proficiency.

The relative magnitudes of the relationship between team performance and the individual performance of the most proficient member of the team on the one hand, and the relationship between team performance and the individual performance of the least proficient member on the other hand, appear to depend upon the nature of the interaction required or permitted by the task. When the group task requires the members to perform parts of the task alternately, Comrey found that team performance tended to have a higher correlation with the performance of the least proficient member of the team. However, when the group task permits greater flexibility of interaction among the members of the team and less structured team task assignments, he found team performance to be more highly correlated with the proficiency of the most proficient member of the team.

The research to be reported here is focused primarily on the second and third of the three questions listed above. However, data relevant to the first question also are noted. The study involves two-person teams working on a cognitive task in which interaction among the members in the performance of the team task is allowed but not specified in its type or amount.

METHOD

Nine jigsaw puzzles of approximately equal difficulty were constructed of ¼" fiberboard. Each puzzle contained 30 pieces and when assembled formed a 1-foot square. The faces of the puzzles were various abstract designs. On the basis of a small standardization sample the puzzles were ranked in terms of difficulty of solution. Being ordered in this manner, the puzzles could be presented in increasing order of difficulty to minimize the effects of practice. The puzzles are considered to be cognitive tasks because they primarily require correct judgments or decisions and only a minimal level of manual dexterity. That

is, the puzzles involve essentially problems of decision making rather than of implementation.

The subjects were 64 pairs of male undergraduates drawn from psychology courses. Most pairs were composed of men who were total strangers, but a number were casual acquaintances and two pairs consisted of men who were friends.

Four puzzles were used to ascertain each subject's level of proficiency on the task while working individually. The puzzles were briefly described to the subject, and the pieces were placed face up in random order on the table before him. The score for a puzzle was the time in seconds required to assemble it completely. One subject in each pair was given the four puzzles, one at a time, to assemble by himself while his partner filled out a personal data questionnaire. After each subject completed his first individual task he exchanged positions with his partner. When both subjects had completed their individual tasks they were given the remaining five puzzles, one at a time, to assemble as a team. The subjects were allowed to talk with one another as much as they wished in deciding how best to proceed as a team. The behavior and the comments of most of the subjects indicated that they enjoyed the tasks and were highly motivated to perform well.

Results

For each individual subject a time score was determined by averaging the times required by him to complete the first four puzzles. Reliability coefficients were estimated from the coefficient of correlation between the scores on the odd and even puzzles corrected by the Spearman-Brown formula. The reliability coefficient for the faster subject was .85, for the slower .90, and for all subjects together .93.

The score for each team was computed as the average time required by the pair of subjects working together to complete the last five puzzles. The reliability coefficient for the team scores was .88. This was obtained by the use of the formula given by Horst (1951) for the situation where the two parts of a test are of unequal length.

An index was computed for each group measuring the extent to which team performance was superior or inferior to the performance expected on the basis of an additive combination of individual skills. This index, designated $D_{(o-p)}$, is the difference between the observed and the predicted team time, and therefore is a measure of team efficiency. The predicted team score was arrived at by determining the amount of time it would have taken the team to finish the puzzles had both members worked at the same rates they did on the individual tasks. If there were no interaction between the members of a team, this index, $D_{(o-p)}$, would be zero: that is, there would be no difference between the team score predicted on the basis of the subjects' performance as individuals and their actual observed team score. A positive value of $D_{(o-p)}$ indicates a negative or inhibitory interaction between the subjects when working as a team, meaning that the team took more time than expected. A negative value of $D_{(o-p)}$ indicates a positive or facilitative interaction between the subjects, meaning that the team took less time than expected. The mean value of $D_{(o-p)}$ for the 64 groups was 18.3, which was significantly different from zero at less than the .001 level. This indicates that on the present task team performance was poorer than would have been predicted on the basis of the pro-

TABLE 1

Means, Standard Deviations, and Intercorrelations among the Various Scores

	Slow	Difference	Sum	Team performance	Team efficiency $D_{(o-p)}$	*M*	*SD*
Fast	.284*	−.164	.605***	.724***	−.219	283.2	72.79
Slow		.898***	.936***	.628***	−.146	469.8	164.48
Difference			.685***	.316**	−.049	186.6	159.99
Sum				.790***	−.200	753.0	197.96
Team					.296*	189.9	40.72
$D_{(o-p)}$						18.3	20.97

* $p < .05$.
** $p < .01$.
*** $p < .001$.

ficiency demonstrated by the members of teams when working individually. Only 10 of the 64 pairs achieved a negative $D_{(o-p)}$ score, that is, did better than expected. Hence with the present task the team situation produced an interfering kind of interaction for most pairs of subjects.

Members of each team were designated as "fast" or "slow" in terms of the speed with which they performed their individual tasks. Table 1 shows that the correlations between the individual times of both types of subjects and team performance are quite high, the correlation in the case of the fast subjects being higher than that of the slow subjects.

The time scores of the two members of each team were added together to yield a "sum" score, and the fast subject's time was subtracted from the slow subject's time to yield a "difference" score. As shown in Table 1, the sum is more highly related to the team score than is either the fast or the slow subject's individual score. Even so, the difference score is significantly related to the team score, indicating that the greater the difference between the subjects the longer it takes them as a team to complete the puzzles. This relationship is to be expected, since the greater the difference between the two subjects the larger is the sum of their performance (Table 1).

That the slow and fast subjects' times are not independent predictors of the team time is indicated by the significant positive correlation between their individual scores. As Comrey (1953) points out, this correlation can be ascribed to the fact that the fast subject's score, the smaller numerical value, cannot be greater than the slow subject's score, the larger numerical value. This imposes a systematic linkage between the individual scores of the pair, thus producing an artifactual correlation.

As may be seen in Table 1, the scores of the fast subjects bear a low negative correlation with the difference scores. Such a relationship is not surprising when it is recalled that all scores are time scores. This negative correlation means that the poorer the fast subject's score the less difference there is between his score and that of his partner. The high positive correlation between the slow subject's scores and the difference scores in part reflects the fact that the variance of the scores of the slow subjects is greater than the variance of the fast subjects. This, too, is a result of placing the fast members of the pairs in one category and their partners in the other. While a slow subject's time score could not be less than his partner's, potentially it could be much larger. Thus, it is seen that the greater the time score of the slow subject, the greater on the average is the difference between the fast and the slow subjects' scores. In other words, the poorer the slow subject is, the less similar to his partner is he likely to be.

The relationships between the index of team efficiency, $D_{(o-p)}$, and all other variables as shown in Table 1 are quite low. While the negative correlations between $D_{(o-p)}$ and the scores of the fast subjects, the slow subjects, and the sum of the scores of both subjects are not significant, they suggest that the more highly proficient a person is, that is, the smaller is his time score, the more difficult it is for him to improve his or his partner's performance in a team situation. The significant positive correlation between team time and $D_{(o-p)}$ scores indicates that the better the team performance, that is the less time it takes the team to solve the puzzles, the more likely it is to perform better than the level predicted on the basis of individual proficiency. This positive correlation is not contradictory to the negative correlations between $D_{(o-p)}$ and individual time scores. It merely emphasizes the fact that fast teams are fast in part because the members are cooperating and interacting effectively, and not solely because the members of the team happen to be highly proficient individually.

The scores and measures of individual proficiency can be combined in at least two different ways in the prediction of team performance. First, the individual scores of the fast and slow subjects in each pair can be used in a multiple regression equation. Secondly, the average or sum of the individual fast and slow scores can be used as one variable and the difference between these scores as a second variable, in another multiple regression equation. The first such equation would deal with the members of the group as individuals, and the second would deal

TABLE 2

COEFFICIENTS OF MULTIPLE CORRELATION AND BETA WEIGHTS FOR PREDICTION OF TEAM PERFORMANCE AND TEAM EFFICIENCY SCORES

	Team performance	Team efficiency $D_{(o-p)}$ scores
Prediction from individual scores of fast and slow subjects	$R = .847$ $\beta_{fast} = .594$ $\beta_{slow} = .459$	$R = .236$ $\beta_{fast} = -.193$ $\beta_{slow} = -.091$
Prediction from the sum of and the difference between the individual scores of the fast and slow subjects	$R = .850$ $\beta_{sum} = 1.080$ $\beta_{difference} = -.425$	$R = .236$ $\beta_{sum} = -.314$ $\beta_{difference} = .166$

with them as a group. These two multiple regression equations are presented in Table 2, where both team performance and team efficiency ($D_{(o-p)}$) are the predicted variables.

Table 2 shows first that the multiple coefficients of correlation using individual scores to predict team performance and team efficiency are only slightly higher than the comparable first-order coefficients presented in Table 1. The multiple coefficients of correlation are .847 for team performance and .236 for team efficiency, while the corresponding first-order coefficients using the sum of individual scores (Table 1) are .790 and −.200. The beta coefficients show that the proficiency of the fast member contributes somewhat more to the prediction of team performance and team efficiency than does the proficiency of the slow member in the multiple regression equations.

The multiple coefficients of correlation in Table 2 that utilize group characteristics, the sum and difference scores, are the same as those that involve individual characteristics. The former coefficients are .850 and .236 and the latter are .847 and .236. The beta weights in the correlations involving group characteristics show that the differences between the members in their individual proficiency contribute less to the multiple predictions than do the averages or sums of the individual scores. With regard to the use of the difference scores in the prediction of the two dependent variables in Table 2, it is especially interesting to note that a relatively large difference between the individuals in their performance as individuals will mean relatively good team performance but relatively poor team efficiency. The situation with regard to the sums, however, is exactly the reverse. Large sums indicate relatively poor team performance but relatively good team efficiency.

DISCUSSION

If the questions raised at the beginning of this article are now reconsidered, it is possible to formulate some answers based on the obtained results in the present study. First, although it was not the primary aim of this investigation, the results support those obtained in other studies that team performance is not as good as performance that would be predicted from an additive combination of individual scores. The task of working together to produce a completed jigsaw puzzle is thus one of a variety of tasks that seems to produce inhibitory effects when two individuals are performing together as a team compared to when they are working apart as two separate individuals.

In answer to the second question of how well team performance can be predicted from a knowledge of the proficiency of individual members of the team, this study showed that relatively good prediction can be obtained from such information. These results are in accordance with those obtained previously by Comrey and his associates on certain of their tasks. However, if the dependent variable is team efficiency (defined as the relation of observed to predicted behavior) rather than team performance, the results obtained in the present investigation show that relatively poor prediction is obtained from measures of individual proficiency. In fact, in the present study there were no variables (neither separate measures of individual proficiency nor their various combinations) that showed any very high correlation with team efficiency. Since this is a measure of team behavior that has not been used in previous similar studies, it is difficult to ascribe reasons for the low predictability of team efficiency or to define those conditions under which such efficiency could be more highly predicted.

A third question posed in the introduction concerned the relative predictability of team

performance by the more proficient member of each team vs. the less proficient member. The findings of the present study show that the scores of the former type of individual, the more proficient member of the team, were slightly more predictive of both team performance and team efficiency. The fact that team performance was somewhat more predictable from the faster member's individual score is in line with Comrey's findings for his cognitive task situation (Comrey & Staats, 1955), but differs from his findings for manual dexterity task situations. An analysis of the procedures used by Comrey in his experiments and those used in the present investigation indicates that the major factor determining whether the faster or slower member's indidvuial score will be more predictive of team performance is not a manual task vs. cognitive task difference, but rather the amount and type of interaction allowed between individuals when performing as a team. For example, in both Comrey's cognitive task procedure and the procedure used here, the faster member was allowed to do more than 50% of the work if he wished; in Comrey's manual dexterity task situations, however, the faster member had to proceed alternately with the slower member, and thus could use his speed on only 50% of the task. Another example of the role of the task determined amount and type of interaction is that in both Comrey's cognitive task and the present task situation the performance acts of one individual provide a variety of cues to the partner that may influence the speed of his performance. As an illustration, the placement of a key piece in the partially completed jigsaw puzzle will give cues as to which two or three other pieces can be added next. Likewise, in the crossword puzzle task, the completion of a vertical word will provide several cues to the partner assigned to the horizontal words. On the other hand, in a task situation where there is simple alternation of performance acts and rigid task assignments, the only cue provided by a single act is the cue for the next specified one to take place.

In the present study, prediction of team performance and team efficiency were computed not only from scores of fast and slow individuals, but also from combination scores, namely, sums and differences of the two individual scores for each team. The sum, of course, significantly predicted team performance, but the difference also was a significant predictor of this criterion. This latter finding regarding the prediction provided by the difference scores is in contrast with Comrey's findings in his manual dexterity task situations. However, even in the present study, the difference scores still offered considerably less prediction of team performance than did the sum scores. Both types of combinations had low correlations with the team efficiency measure, although again, the sum scores produced somewhat higher predictability.

Summary

The present investigation examined the relationship between the proficiency of subjects solving a cognitive-type task individually and in teams of two persons. It was found that team work was on the average less productive than the sum of individual performances. Team performance was found to be better predicted by the individual proficiency of the better member than by the proficiency of the poorer member. A simple additive combination of the individual scores of both members of the teams gave a good prediction of team performance, and this prediction was only slightly increased when the individual proficiencies of the two members were optimally weighted in a multiple regression equation. The more similar the two members of a team were in individual proficiency, the more likely they were to form a proficient and effective team. The extent to which the two members of a team facilitated or interfered with each other when working together was predicted quite poorly from measures of individual performance. Several of the above results were discussed in terms of the amount and type of interaction allowed between the two individuals when performing together as a team.

REFERENCES

Comrey, A. L. Group performance in a manual dexterity task. *J. appl. Psychol.*, 1953, **37**, 207–210.

Comrey, A. L., & Deskin, G. Further results on group manual dexterity in men. *J. appl. Psychol.*, 1954, **38**, 116–118. (a)

COMREY, A. L., & DESKIN, G. Group manual dexterity in women. *J. appl. Psychol.*, 1954, **38**, 178–180. (b)

COMREY, A. L., & STAATS, CAROLYN K. Group performance in a cognitive task. *J. appl. Psychol.*, 1955, **39**, 354–356.

HORST, P. Estimating total test reliability from parts of unequal length. *Educ. psychol. Measmt.*, 1951, **11**, 368–371.

HUSBAND, R. W. Cooperative versus solitary problem solution. *J. soc. Psychol.*, 1940, **11**, 405–409.

LORGE, J., FOX, D., DAVITZ, J., & BRENNER, M. A survey of studies contrasting the quality of group performance and individual performance: 1920–1957. *Psychol. Bull.*, 1958, **55**, 337–372.

MARQUART, DOROTHY I. Group problem solving. *J. soc. Psychol.*, 1955, **41**, 103–113.

SHAW, M. E. Comparison of individuals and small groups in the rational solution of complex problems. *Amer. J. Psychol.*, 1932, **44**, 491–504.

TAYLOR, D. W., & FAUST, W. L. Twenty questions: Efficiency in problem solving as a function of size of group. *J. exp. Psychol.*, 1952, **44**, 360–368.

WATSON, G. B. Do groups think more efficiently than individuals? *J. abnorm. soc. Psychol.*, 1928, **23**. 328–336.

(Received February 3, 1961)

JUST PUBLISHED . . .

TESTS IN PRINT

A COMPREHENSIVE BIBLIOGRAPHY OF TESTS FOR USE IN EDUCATION, PSYCHOLOGY, AND INDUSTRY

Edited by

OSCAR K. BUROS

Director, The Institute of Mental Measurements
Rutgers, The State University

A valuable new reference work for test users and students of testing

- **Concise information** on more than 2,000 tests published as separates in English-speaking countries.
- **Comprehensive coverage**: achievement batteries and tests in specific areas, projective and non-projective personality tests, intelligence and scholastic aptitude tests, multi-aptitude batteries, vocational interest, aptitude, and selection tests, and numerous tests in other areas are presented.
- **Complete cross references** to additional information, critical reviews, and test references to be found in any of the volumes of *The Mental Measurements Yearbook* series published to date. Serves as a classified index and supplement to the yearbook series.
- **Brief information** (title and yearbook cross references only) on more than 800 out of print tests for the convenient reference of yearbook users and students of testing.
- Other materials of interest to test users.
- Complete, convenient indexes.

Order directly from Pp. xxix, 479 $7.00

THE GRYPHON PRESS 220 Montgomery Street Highland Park, New Jersey

ANNOUNCING

Devoted to the Technology of Human Learning

Thomas F. Gilbert, EDITOR IN CHIEF
Carroll Dunn, EXECUTIVE EDITOR

EDITORIAL ADVISORS

John A. Barlow
Joseph V. Brady
Charles B. Ferster
Israel Goldiamond
Paul S. Siegel
Charles W. Slack

Papers of profound interest to psychologists, educators and human engineers on new applications of reinforcement are becoming available in growing numbers.

The Journal of Mathetics fills the need for an organ devoted to their collection.

It gives the scientists in this field a publication of professional standing in which they may discuss their own work and keep informed on the findings of their contemporaries.

Published bi-monthly from January 1962

Annual subscription $18. A.P.A. members $12.
A.S.T.D. members $12. Students $10.

THE JOURNAL OF MATHETICS
Box 3232-A, University, Alabama